

Going by the book

In Italy's election campaign, opposition parties have pledged research reform — but nothing will change until agency chiefs start playing by the rules.

If Dante were writing his *Divine Comedy* in today's Italy, he might still find cause to pen the line: "Laws do exist, but who is there to apply them?" But the inspiration this time would not be encounters in Purgatory with the rich and famous of medieval Italy. Instead, he might imagine meeting some of Italy's leading university professors and the presidents of its research agencies. These tortured souls would recall the many and detailed laws governing Italian academia, and their own failure to apply them.

For decades, Italy's scientists have worked under regimes that distribute money and academic positions without much fairness or transparency. Attempts to create a more equitable system have stumbled repeatedly. Just trying to survive in the system requires lots of energy that might otherwise be devoted to good science. Money has been in short supply for basic research in Italy for years, and fell after the election of Silvio Berlusconi's government in 2001. But even this cannot fully explain the feeling of stagnation in Italian research.

Rather, it is the failure to implement reforms that most depresses Italian science. For example, laws already exist that would permit researchers at the institutes of the CNR, Italy's main basic-research agency, to apply for promotion. The CNR, whose president is Fabio Pistella, is supposed to hold competitions for internal promotions every year or two. But the last one was completed seven years ago. The most recent competition was opened in 2004, yet the evaluation committee was appointed just a few weeks ago and scientists don't expect results any time soon.

There are laws allowing Italy's space agency, the ASI, to carry out space-science research, one of Italy's strengths — in fact, such research is a major part of the ASI's official mission, and the agency is relatively flush with money. Yet under the leadership of Sergio Vetrilla, it has not started a new national research programme for nearly five years.

These two examples of playing safe by doing nothing are emblem-

atic of a system in which no one has to bear personal responsibility and take the consequences.

This culture has to change. Romano Prodi, leader of the centre-left coalition that is ahead in the polls in the run up to the election on 9 April, has been listening to the woes of Italy's scientists (see page 264) and has promised to turn things around if he wins. His science advisers are aware of the problems and pledge that agency chiefs would be selected by Prodi from those competent and willing to make appropriate decisions, and that systems would be established at universities to render administrators accountable for their actions.

This could mean, for example, that when an academic appointment is made, the dean or rector involved would no longer be able to duck the blame if the professor proved to be an unproductive researcher. At the moment the buck can be passed widely, thanks to, among other things, the requirement in universities for secret faculty votes, or the fact that a candidate has been preselected by others onto the new national list of those deemed qualified to be a professor. The sort of independent evaluation that Prodi's advisers are advocating would attach consequences to bad decisions. Universities with poor research records would get less government money, and might hold their own rectors or deans accountable for the fact.

"The sort of independent evaluation that Prodi's advisers are advocating would attach consequences to bad decisions."

If such a system of personal responsibility sounds familiar, it is because it is already well established in most other scientifically advanced nations. But cultural change won't occur overnight in Italy. The structural adjustments that Prodi is being advised to make could nurture it, however. This would increase the chance of Dante encountering academic power-brokers not in the circles of Hell or the corridors of Purgatory, but in the meadows of Paradise. ■

Dreams of flu data

The lack of an accessible store of information is undermining the fight against avian flu.

"Confidentiality of sensitive national outbreak surveillance data assured!" This prominent guarantee on the website of the South East Asian Nations Infectious Diseases Outbreak Surveillance Network says it all. Open sharing of data often ends when it could compromise trade or other national interests.

Massive public databases exist in many areas of science, and are critical to cutting-edge research. But there is no comprehensive database of outbreaks of infectious diseases. We have better data on

galaxies 10 billion light years away than on human cases of avian flu in China or Vietnam. Yet the world is imperilled by outbreaks, wherever they happen.

It is difficult enough to gather the data in the first place. At the moment, plague, cholera and yellow fever are the only notifiable diseases that countries must report to the World Health Organization (WHO). Vietnam has often reported human cases of avian flu months after the event, and outbreaks in animals have been concealed in many countries.

Not before time, the WHO will have broader powers in 2007, when international health regulations, agreed by its members in May last year, come into force. These impose obligations on states to respond to any infectious disease of international concern. Cat-and-mouse games will no doubt continue, but the WHO will at



least have a 'health policing' role, something that it currently lacks.

The reporting of avian-flu cases has recently improved in speed and openness, but the quality of the available data remains dire, and biological samples are insufficient. A typical WHO update will give the number of new cases from a country and, on a good day, the age, sex and rough location of each case. But there is little information on familial case clusters, and typically no clinical data. What is worse, these few data are in text form strewn across hundreds of individual WHO web pages.

Data on outbreaks in poultry are even more sparse, and mostly come from the World Organisation for Animal Health (OIE). Someone at the UN Food and Agriculture Organization (FAO) is maintaining a file of cumulative bird outbreaks from OIE and other data, and is making it available to researchers and journalists. But it is incomplete, lacks good location data and contains errors.

Genetic data are also lacking. When samples are sequenced, the results are usually either restricted by governments or kept private to an old-boy network of researchers linked to the WHO, the US Centers for Disease Control and Prevention, and the FAO. This is a far cry from the Human Genome Project, in which all the data were placed in the public domain 24 hours after sequencing. Many scientists and organizations are also hoarding sequence data, often for years, so they can be the first to publish in academic journals. With the world facing a possible pandemic, such practices are wholly unacceptable. *Nature* and its associated journals are not alone in supporting the rapid prior exposure of data when there are acute public-health necessities.

Three cheers, then, to Ilaria Capua of the Tri-Veneto Region Experimental Animal Health Care Institute in Italy, who last month threw down the gauntlet to her colleagues by refusing to put her latest data on Nigeria and Italy in these private networks. Instead she uploaded them to GenBank and called on her colleagues worldwide to do likewise. Only in this way can researchers establish and track the global pattern of the evolution of the bird-flu virus.

Imagine scientists anywhere being able to log on to a publicly available, searchable Internet database, updated in real time, with full clinical and sequence data on each human case, and accurate and complete poultry

"The world badly needs a database for outbreaks of avian flu."

data. Dream on. The WHO's clunky online *Global Health Atlas*, which gives rough aggregate data for many diseases, doesn't have a category for H5N1 under influenza. The Global Infectious Diseases Epidemiology Online Network (GIDEON) database contains only the data on avian flu that it extracts from the WHO's updates and reports from the ProMED reporting system for infectious diseases. ProMED itself has pioneered outbreak alerting, notably during the SARS crisis, but its content consists largely of media cuttings. Such aggregation is often done faster and better by bloggers, and by the Global Public Health Intelligence Network (GPHIN), a Canadian intelligence operation that provides an early warning system that screens media and blogs in seven languages in real time.

The world badly needs a database for outbreaks of avian flu. But international agencies are failing to rise to the challenge. ■

The right chemistry

Nature celebrates a discipline's unheralded achievements.

Television and cinema aren't often kind to chemists, who regularly find themselves portrayed as the nefarious creators of toxic pollutants, or as mad scientists brewing up Love Potion #9 in some cluttered and archaic laboratory.

But most chemists innocently pass their time trying to figure out how things work at the molecular level, often using a relatively simple set of concepts to shed light on complex natural phenomena.

Fireflies, for example, communicate with each other by emitting light, and the protein responsible for this bioluminescence reaction is luciferase, which is well known to biologists. In this issue of *Nature*, Nakatsu and co-workers explain how they used synthetic chemistry, structural biology and biochemistry to explore how changes in the active site of the protein lead to changes in the colour of the emitted light (see pages 285 and 372).

The work is published as part of an issue in which we have collected, ahead of the American Chemical Society's meeting in Atlanta, Georgia, several chemistry and biochemistry papers along with an interview with Nobel laureate Roald Hoffmann on his work with young scientists in the Middle East (see page 274) and a *NatureJobs* assessment of career opportunities in green chemistry (see page 378).

In addition, *Nature's* website now includes a collection of recently

published *Nature* papers on metalloproteins, a set of proteins containing transition metals that are involved in a wide range of biologically significant processes, including DNA repair and the biosynthesis of natural products (see www.nature.com/nature/focus/metalloproteins).

Our website also features a chemistry blog launched this week, 'The Sceptical Chymist', featuring entries by editors at *Nature* journals as well as by authors and readers (see <http://blogs.nature.com/thescepticalchymist>). The first blog entry explains the choice of title.

Last, but not least, the website will feature a podcast including lively interviews about chemistry and biochemistry papers recently published in *Nature* and its sister journals. The podcast and an online collection of recent chemistry content will be available from 23 March (see www.nature.com/conferences/acs).

"We hope that this modest barrage of publishing activity will help to convey some of the excitement to the wider scientific community."

Chemists themselves probably don't need to be reminded of the usefulness of their work, or of the excitement inherent in it. We hope, however, that this modest barrage of publishing activity will help to convey some of that excitement to the wider scientific community — and even to interested observers in the world beyond. Perhaps one day even Hollywood producers will recognize the diligence with which chemists attempt to interrogate nature, and the inherent value of their contribution to our understanding of it. But don't hold your breath. ■

RESEARCH HIGHLIGHTS

Go with the airflow

Geophys. Res. Lett. **33**, L04804 (2006)
The flow of warm, dry air from the Sahara to the Atlantic Ocean is thought to inhibit the formation of tropical storms. Now, NASA researchers have shown that careful modelling of this airflow, known as the Saharan air layer, can improve simulations of storms forming off the coast of western Africa. Liguang Wu of the Goddard Earth and Technology Center in Baltimore, Maryland, and his team used satellite data to build atmospheric temperature and humidity profiles. They then simulated the formation of Hurricane Isabel (pictured), which hit the United States in 2003, with and without this detailed data describing the Saharan air layer. Inclusion of the data yielded a much more accurate model of Isabel's observed track.



M. TRENCHARD/JOHNSON SPACE CENTER

OPTICS

Mirror image

Appl. Phys. Lett. **88**, 091119 (2006)
Mirrors don't just reverse right and left — they also invert the phase of light waves that bounce off them. More precisely, they reverse the phase of the electrical component of an electromagnetic wave, but leave the magnetic component intact. Nikolay Zheludev of the University of Southampton, UK, and his colleagues have now made a mirror that does the opposite. Dubbed a 'magnetic wall', the reflector preserves the electrical phase of reflected microwaves, but reverses the magnetic phase. It is a metamaterial made from undulating copper strips 0.8 millimetres wide and arranged in a 'fish-scale' pattern.

CELL BIOLOGY

Fever pitch

J. Neurosci. **26**, 2590–2597 (2006)
A problem in protein transport could underlie some cases of febrile seizures, the fever-triggered convulsions that afflict over 6% of children worldwide. Febrile seizures have been linked to mutations in one subunit of the GABA_A receptor, a protein complex that is a key ion channel in the nervous system. Robert Macdonald of the Vanderbilt University Medical Center in Nashville, Tennessee, and his colleagues show that the amount of the GABA_A receptor at the surface of mammalian cells drops abnormally within minutes

of the thermometer jumping from body temperature (37 °C) to a feverish 40 °C. This could be how a rapid rise in body temperature triggers the seizures, the authors propose.

CHEMICAL BIOLOGY

Small yet perfectly formed

Nature Chem. Biol. doi:10.1038/nchembio775 (2006)
Hard to find, but valuable: this is the nature of a small molecule that specifically binds to just one protein. Such molecules can be used to elucidate the cellular function of the protein, or in therapeutic treatments. Thus Eric Prossnitz of the University of New Mexico in Albuquerque and his co-workers are set to reap the rewards for their hard work. The team has discovered the first small molecule that binds to the oestrogen-binding G-protein-coupled

receptor GPR30, but not to the two oestrogen receptors ER α and ER β (crystals of oestrogen pictured below). They used computational methods and cell-based assays to find the molecule, which they hope to use to define the role of GPR30 *in vivo* and to develop new contraceptive and anti-cancer agents.

MICROFLUIDICS

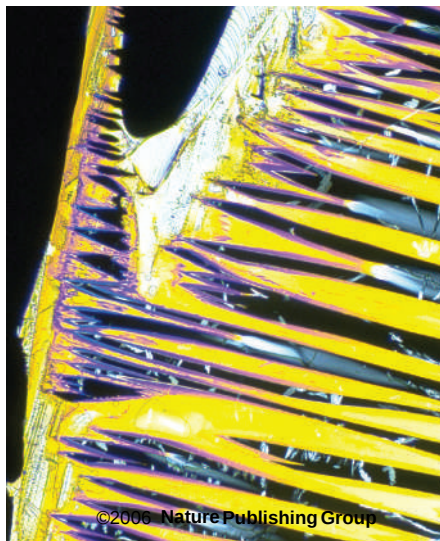
In a heartbeat

Lab Chip **6**, 362–368 (2006)
Researchers in Japan looking for a new way to power microdevices have turned to one of nature's most dynamic inventions: the heart. A group led by the University of Tokyo's Takehiko Kitamori tethered a sheet of rat cardiomyocytes to a microfluidic chip. It is the contraction and expansion of layers of these cells that makes the heart beat. On the chip, the cardiomyocyte sheet is used as a pump actuator. It generates enough force, at a few microneutons per cell, to drive fluids in the chip's channels. The cells are powered chemically, with glucose and oxygen. But Kitamori says that to be viable in applications such as drug delivery, the bioactuator needs to be longer lasting.

CELL BIOLOGY

Saved by cleavage

Nature Cell Biol. doi:10.1038/ncb1378 (2006)
One of the cell's mechanisms for coping with DNA damage is regulated by a self-destructing enzyme, say researchers. Alan D'Andrea of the Dana Farber Cancer



OXFORD SCIENTIFIC (OSF)

Institute in Boston, Massachusetts, and his colleagues focused on a mechanism that exploits the protein PCNA to prevent DNA replication stalling at a point of damage. When a ubiquitin molecule is attached to PCNA, damage-tolerant DNA polymerases are recruited — but this is risky because the polymerases are not faithful replicators.

D'Andrea's group identified an enzyme, USP1, that curbs the risk when the chances of damage are low by stripping ubiquitin from the protein. They also showed that USP1 can bow out of the system, leaving PCNA ubiquitinated, when there is a strong possibility of DNA damage. Irradiation with DNA-damaging ultraviolet light caused USP1 to cleave, and therefore inactivate, itself.

ASTRONOMY

Spin cycle

Astrophys. J. **639**, 1238–1251 (2006)

Discovered among the icy debris at the outskirts of the Solar System, the Kuiper belt object 2003 EL61 is the fastest-spinning body in our neighbourhood that is greater than 100 kilometres across. So say David Rabinowitz of Yale University in New Haven, Connecticut, and his colleagues.

Each rotation of the body takes less than four hours, and the authors believe that this rapid spinning squeezes the icy object into an ellipsoid at least 1,960 kilometres long and as little as 500 kilometres wide. After Pluto and 2005 FY9, 2003 EL61 is the third brightest trans-Neptunian object — orbiting the Sun at a greater distance, on average, than Neptune. Its light spectrum suggests that its surface is covered in methane or water ices.

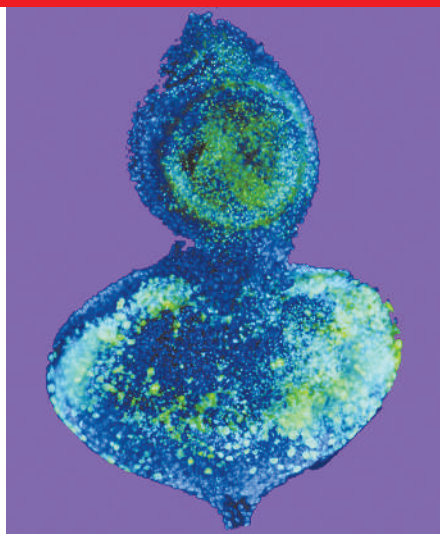
DEVELOPMENTAL BIOLOGY

Telling up from down

Cell **124**, 1011–1023; 1025–1037 (2006)

Cells need to know which way is up to coordinate growth, division and migration, and biologists are curious to know how they make this distinction.

Two groups of researchers, led by Christine Jacobs-Wagner of Yale University in New Haven, Connecticut, and Patrick Viollier at Case Western Reserve University in Cleveland, Ohio, studied the process in the crescent-shaped bacterium *Caulobacter crescentus*, which divides asymmetrically to create two different daughter cells. The teams show that a protein they dub 'tip of new pole' (TipN) marks one pole of the dividing cell, and then moves to the new poles of each progeny cell. TipN works by directing construction of a flagellum at one end of the cell.



CANCER

Location, location, location

Dev. Cell **10**, 303–315 (2006)

Abnormal chromosome structure is a hallmark of cancer and certain birth defects. Now, researchers have uncovered a mechanism that could trigger such problems.

Gary Karpen of the Lawrence Berkeley National Laboratory, California, and his colleagues studied a protein called CID that helps to form the kinetochore, a structure needed for chromosomes to be pulled into daughter cells during cell division. Each chromosome normally has one kinetochore. Working in the fruitfly *Drosophila melanogaster* (eye-antenna disc pictured above), the team found that abnormally high levels of CID (shown in green) resulted in the protein sticking to extra locations on the chromosomes, causing additional kinetochores to form. The chromosomes were then pulled to pieces during cell division.

COMPUTATIONAL CHEMISTRY

Molecular map-making

J. Am. Chem. Soc. **128**, 3228–3232 (2006)

Designing a molecule with tailor-made properties is laborious when it involves testing many possible structures. One way to speed this process up, suggest David Beratan, Weitao Yang and their colleagues at Duke University in Durham, North Carolina, is to convert a discrete set of combinations of atoms into a continuous 'molecular space' in which atoms — and their contributions to the property of interest — can be gradually turned 'on' and 'off'. This smooth terrain can be explored with well developed mathematical tools to search for peaks, corresponding to maxima of the desired property. The nearest positions to these peaks that correspond to discrete molecules represent the candidates of interest.

The researchers used the technique to find the optimal electronic polarizabilities for simple small molecules; but the generality of the approach remains to be seen.

JOURNAL CLUB

David MacMillan
California Institute of
Technology, Pasadena

**Bond with a chemist over
the search for asymmetry
in reactions.**

In synthetic chemistry, even the conceptually simple task of creating a chemical bond can be a formidable challenge.

Symmetry in particular can be problematic. Many organic molecules are chiral; that is, not superimposable on their mirror image. The configurations in which such a molecule exists, known as enantiomers, often differ in behaviour, especially in biological systems. So chemists strive for ways to preferentially produce just one enantiomer.

The design of such 'enantio-selective' reactions is central to synthetic chemistry. Significant advances have been made using new catalysis concepts, but many key problems remain unsolved.

A challenge I had always thought particularly daunting was developing a catalytic system to join an electrophilic alkyl group to a nucleophilic carbon centre in a second alkyl group, with control over the symmetry of the product.

There are two complicating issues. First, the alkyl-transition metal complexes mostly used as the starting material are unstable and so decay into metal hydrides. Second, the alkyl halides that act as reactants are themselves chiral and are a racemic, or 50:50, mixture of enantiomers. Thus, we must engineer a process that simultaneously destroys chiral information in the reactants and creates asymmetry in the product.

I was excited to learn that this is now possible. Gregory Fu and his colleagues have recently described efficient nickel-catalysed couplings of a racemic alkyl halide with organozinc nucleophiles, and they achieve superb enantiocontrol (F. O. Arp and G. C. Fu *J. Am. Chem. Soc.* **127**, 10482–10483; 2005). The method is ground-breaking as a proof of concept, and I predict that it will be widely used.

NEWS



Dusty answer: these blocks of aerogel contain grains from a comet's tail.

NASA

Comet chasers get mineral shock

LEAGUE CITY, TEXAS

The first results from a mission to catch dust from a comet's tail have revealed a surprise: these balls of dirty snow are born of fire as well as ice. Scientists were stunned to find a huge range of minerals in the particles captured by NASA's Stardust probe as it swooped past the comet Wild 2 on 2 January 2004. Many of the compounds could only have formed close to a star — far from the chilly outskirts of the Solar System where the comet first coalesced.

The Stardust team presented its preliminary results to a packed room of more than 600 scientists at the Lunar and Planetary Science Conference in League City, Texas, on 13 March. The grains are the first material ever brought back from a comet (see 'Caught in the wild'). Indeed, they are the first geological samples collected from space since the Soviet Luna 24 mission brought back Moon rocks two decades ago.

Of the 30 or so microscopic grains studied so far, almost a quarter are of a type known as calcium–aluminium inclusions (CAIs). These solidify at several thousand degrees and have never been detected in remote observations of comets. "When these grains first formed they were white hot," says Don Brownlee of the University of Washington, Seattle, who is the mission's chief scientist.

One explanation for the grains' presence is that they were created very close to the young Sun, and then flung to the outer reaches of the dusty cloud that spawned the planets. If that is the case, planetary scientists have a lot of rethinking to do — it would mean that the primitive Solar System saw a lot more mixing than previously believed.

The presence of CAIs was predicted by one controversial theory, however. The 'X-wind'

model sees strong magnetic fields around the young Sun channelling heated material to the far reaches of the protoplanetary cloud. The Stardust results do not prove anything, but they certainly fit with the idea, says Mike Zolensky of NASA's Johnson Space Center in Houston, Texas, who is leading the mineral analyses.

The other possible source of the minerals is another star completely. Once formed, the grains may have drifted through interstellar space for eons, before reaching our own Solar System and being incorporated into Wild 2. Zolensky says it is too early to know which theory is right. But Brownlee adds that measurements of isotopes in the grains could settle the question within months — such materials often bear a distinctive isotopic signature determined by their star of origin.

Scientists involved in the work are jubilant about the progress so far. Last month, *Nature* spent a day with Phil Bland, a planetary scientist at Imperial College London, as he and his team analysed part of a grain (www.nature.com/news/2006/060213/full/060213-2.html).

Caught in the wild

The 5-kilometre-wide Wild 2 was first seen in 1978 by Swiss astronomer Paul Wild. It was chosen for the Stardust mission because it spent billions of years orbiting the Sun at around the same distance as Uranus. At this distance it would have been relatively unaffected by the Sun's heat, so there is a chance that its chemical make-up has been virtually unchanged since it was born. In 1974, a close encounter with Jupiter spun the comet towards the Sun; it now spends some of its orbit even closer to the Sun than Mars, making it relatively accessible to a visiting spacecraft.

M.P.

Nature was asked by NASA not to report the results at the time, but Bland was excited to find large amounts of calcium. He speculated that it might be present in the form of calcium carbonate — even making a bet with his colleague Matt Genge. This mineral almost always forms in water, and would imply that the comet had been heated enough for its ice to melt, perhaps by collisions with meteorites.

Bland's sample has now been shown to be a CAI, and NASA scientists have not yet found any carbonates in their grains. So Bland owes Genge dinner, but he is still overjoyed to have found hints of the comet's fiery past.

And he is not the only one. "Stardust has exceeded all my expectations," says Peter Tsou of NASA's Jet Propulsion Laboratory in Pasadena, California. Tsou developed the technique of catching comet grains in an aerogel, an extremely lightweight, glass-like substance, and he has been overseeing efforts to extract dust from the carrot-shaped impact tracks. "This has been my dream for 25 years," he says.

Tsou adds that the success of Stardust vindicates the sometimes maligned 'faster, better, cheaper' approach to NASA missions in the 90s, which was designed to focus on low-budget missions over short timescales. Some scientists fear that such missions are now in danger of being overlooked in favour of giant projects such as the James Webb Space Telescope (see *Nature* 440, 140–143; 2006), especially in a NASA budget that is feeling the strain of the shuttle and International Space Station programmes.

In the meantime, work on the grains continues. Thousands of particles have yet to be analysed, says Brownlee, and "we could easily work for a year on a single particle".

Mark Peplow

Signs of warm water on Saturn's moon

Astrobiologists have welcomed the news that Saturn's small moon Enceladus has a huge geyser that spews water vapour and dust thousands of kilometres into space. It means the satellite may be hiding one of the most accessible reservoirs of liquid water in the Solar System.

Evidence for the geyser was discovered by NASA's Cassini probe during three fly-bys in 2005. "Cassini essentially flew through it," says planetary scientist Jonathan Lunine of the University of Arizona, Tucson.

According to a suite of papers in last week's *Science* (see *Science* 311, 1388; 2006), the plume, which also contains carbon dioxide, methane, nitrogen and propane, is probably fed by a reservoir of liquid water that nestles just beneath the icy surface.

The idea that material could be welling up beneath the surface of Enceladus isn't new. When the Voyager probes flew past in 1980 and 1981, they spotted a flat area around the south pole that looked as though it had recently been resurfaced

by volcanic activity; there was no sign of craters left by meteorite impacts, suggesting they had been wiped away. Astronomers also speculated from Voyager data that Saturn's E ring of fine particles was being fed by material from Enceladus.

The latest observations

support that idea, and suggest that the geyser could in part be driven by heat from the movement of Enceladus' insides, which

slosh around as the moon is pulled back and forth by its neighbours' gravity. Heating from the decay of radioactive elements also contributes to the process. The Cassini team estimates that at least 150 kilograms of material a second comes from the plume, which is more than enough to supply Saturn's E ring.

But the probe did not find any ammonia — previously suggested as the antifreeze that helps keep internal water liquid in this chilly part of the Solar System. So parts of

Enceladus may be much warmer than anyone expected, a possibility that has excited astrobiologists.

"We have found another environment in our Solar System, in a very surprising place, that could host living organisms," says Caroline

Porco of the Space Science Institute in Boulder, Colorado, who leads Cassini's imaging team.

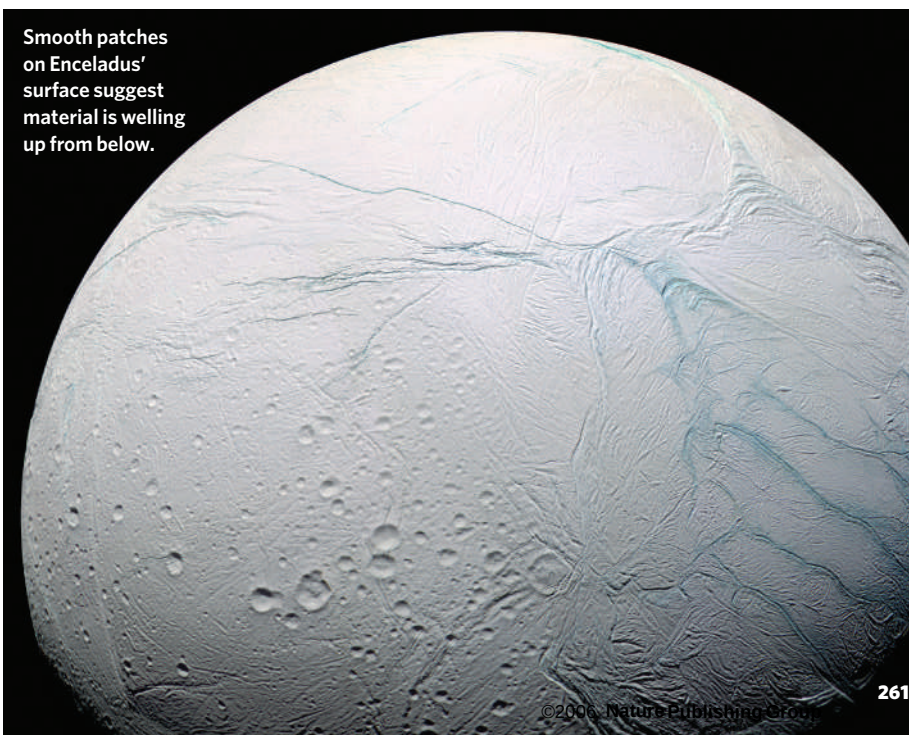
That makes Enceladus an attractive place

to explore, says Lunine. The ice-covered oceans of Jupiter's large moon Europa remains the main target of astrobiologists' interest. But its frozen crust may be kilometres thick. "What's different here is that pockets of liquid water may be no more than tens of metres below the surface," says Andrew Ingersoll, an atmospheric scientist at the California Institute of Technology, Pasadena. Cassini's next close fly-by in spring 2008 should help to find out for sure.

Mark Peplow

"Pockets of liquid water may be no more than tens of metres below the surface."

Smooth patches on Enceladus' surface suggest material is welling up from below.



ON THE RECORD

“How can we lose... when we have a NASA engineer on our team?”

A contestant on the television show *Survivor Panama* is despondent after former astronaut Dan Barry fails in a team challenge. Barry was later evicted from the show by his team.

“It was the most rewarding poster I’ve ever done.”

Geologist Alison Rust muses on her experiments that compared lava flows to fudge.

Sources: *CollectSpace.com*, New York Times

SCORECARD



Genes and architecture
Singapore has hatched a plan to build a major bridge in the shape of the DNA double helix.



Roboshark
US Navy researchers are developing a shark-tracking tag they can implant in the creature’s brain, which might someday control its swimming.



Weather forecasting
There is a new kind of betting on the US national collegiate basketball tournament this month. Correctly predicting the weather on competing campuses on the day of a game could net you a week-long tornado chase.

NUMBER CRUNCH

Ever heard something dodgy on the evening news? A survey of local news broadcasts in 50 US cities backs up the notion that television isn’t always a reliable place to get your health and medical advice.

1,799 health stories were broadcast during 2,795 top-rated news shows during October 2002.

33 seconds was the median length of those segments.

27% included an interview with a health professional.

23 reports described how duct tape could be used to remove warts.

Source: Pribble, J. M. et al. *Am. J. Manag. Care* **12**, 170–176 (2006).

Consumer products leap aboard the nano bandwagon

WASHINGTON DC

The number of commercial products advertised as containing nanoparticles is increasing rapidly, according to a new inventory. Environmental groups say the list shows that not enough is being done to oversee nanotech’s spread into the commercial sphere. But others say that marketing, not a rise in use of the technology, may be driving the trend.

The inventory of nanogoods was released on 10 March by the Woodrow Wilson International Center for Scholars, a think tank based in Washington DC that works on nanotech issues. Cataloguing every nanotech product that exists would be close to impossible — there are no regulations that require companies to register such products. So for a rough estimate of how much nanotech is out there, Andrew Maynard and his colleagues at the centre scanned the web for products that openly advertise the use of nanotechnology.

Maynard and his co-workers found 212 products that use nanotechnology. This is double the number found by a similar survey carried out last year by EmTech Research, a pro-industry research group based in Ann Arbor, Michigan. Nearly half of these products were creams, cosmetics and supplements, designed to be applied to the skin or taken orally.

That is evidence of the industry’s growing commercial success, says Francine Porter, president of Denver-based Osmotics Cosmeceuticals. Her company markets an anti-cellulite cream called Lipoduction, describing “Nano Technology that increases delivery up to 700% over traditional cellulite products.” Porter says sales have been brisk. “This industry is just exploding with nanotechnology,” she says.

Environmental campaigners, who have long voiced fears about the health and environmental implications of nanotechnology, say the figures highlight a worrying lack of regulation. “If the numbers are to be trusted, it says to me that potential exposure to nanomaterials would appear to be growing quicker than expected,” says Douglas Parr, a physical chemist and chief

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Small wonder: advertising the use of nanotech may hook buyers.

scientist for Greenpeace UK in London.

But others say much of the apparent surge in the use of nanotechnology may be the result of companies relabelling their goods to meet consumer preferences. For example, most cosmetic creams already contain nanoscale particles to penetrate the skin, so companies could use this in their marketing. “Non-scientists tend to think there is something magic about nano,” says Jöns Hilborn, a chemist at Uppsala University in Sweden and former president of the European Tissue Engineering Society.

Scientists have already learned to use this relabelling trick to win funding from politicians, says Hilborn. A project he heads, to develop miniature scaffolds for tissue engineering, recently won €1.7 million from the European Union’s Framework programme, following a call for nanobiotechnology projects. “I could have very well written the proposal without nano in there,” he says. “I didn’t lie to get the money; I just used the word they like to hear.”

Paul Ferron, who heads Beyond Skin Science — a California-based company that sells a line of nanotechnology-based products — agrees that despite the concerns of campaigners, the term ‘nano’ is increasingly becoming a selling point for consumers as well as funding agencies. Products such as Apple’s iPod nano music player have boosted recognition of the word, he says. “I see it more and more, I hear it more and more.” ■

Geoff Brumfiel

Additional reporting by Jenny Hogan

GULLIVER/ZEFA/CORBIS



RODENT RISES FROM THE DEAD

Mammal is the living representative of a lineage presumed extinct.

www.nature.com/news

Huge Biobank project launches despite critics

British family doctors have started recruiting subjects for an ambitious effort to probe the interplay of genetic and environmental factors that cause disease.

Researchers hope that half a million people will donate blood and urine samples, and give access to their medical records, as part of UK Biobank. This £61-million (US\$105-million) project launched on 13 March, funded by the Medical Research Council and the Wellcome Trust. Critics question whether it will give value for money. But the project's organizers say it has enormous research potential.

Participants will be weighed and measured, and answer questions on topics such as alcohol intake and family medical history. Details will be held in a database and linked to records collected by the National Health Service (NHS).

Thousands of cases of conditions such as heart disease and cancer are likely to develop,

and researchers will be able to search the database for underlying genes and lifestyle choices.

Critics, including members of the House of Commons Select Committee on Science and Technology, say health records alone can be a poor guide to the conditions a person suffers

"Complaints stem in part from a misunderstanding of the scheme."

from. Medical journals have asked whether enough cases of disease will develop for results to be statistically meaningful. In contrast to some other genetic-database projects, however, critics are not concerned about data-privacy and ethical issues.

UK Biobank's chief executive, Rory Collins, an epidemiologist at the University of Oxford, says complaints stem in part from a misunderstanding of the scheme. It might make sense to study the families of ill people if disease genes were the main target, he says, but they are not.

"The primary use will be to compare people who develop disease with those that don't," he

says. Researchers will then be able to examine the lifestyle and genetic factors that differentiate the two groups.

Collins estimates that scientists will want to start mining the database in about ten years, when several thousand cases of diseases will have emerged. He adds that changes in the NHS computer system should have made it easier for scientists to use for research by then.

Access will be controlled by UK Biobank, which will vet applications and coordinate groups to minimize the cost of accessing samples. These will be frozen and split between two locations, to provide a back-up.

The UK project will be tracked in countries with similar schemes. Canada, Iceland and Japan have set up smaller biobanks. Work began in 2001 on a project in Estonia, where the government wants to collect genealogical, health and DNA data on a million citizens — it aims to have data on some 100,000 by 2009. ■

Jim Giles

SPECIAL REPORT

Saving Italian science

As the general election looms, candidate prime minister Romano Prodi strives to convince Italy's discontented scientists that he can turn things around. **Alison Abbott** reports.

Last autumn, Romano Prodi, a candidate for Italian prime minister, proposed a thought experiment to a group of top scientists. "If you had an additional €400 million (US\$480 million) a year for five years to rescue Italian research, to be allocated during the first 100 days of government, what would you do with it?" he asked.

Prodi had summoned 20 or so scientists to his *Fabbrica*, a think-tank housed in Bologna where he has been developing his government platform.

The short answer came fast: "We would double the number of researchers." Italy's research force is currently half the size of comparably large, rich countries. The longer answer comprised a ten-page document published last December, which lists the many problems with Italy's underperforming research sector, and how they might be tackled. The bottom line: too much bureaucratic incompetence and an unreasonable demand for immediate returns, as well as too little money and meritocracy.

The conclusions won't surprise most Italian scientists. "Physicists have developed a theory for chaos, but Italy is now running an experiment in chaos," comments Carlo Rubbia, a 1984 physics Nobel laureate and, until last year, president of the Italian energy and environmental agency ENEA.

But the document proposes how to bring order. Its concerns and recommendations are now feeding back into detailed action plans for a government science programme that can be agreed by all nine parties of Prodi's centre-left Olive Tree coalition.

Mathematician Luciano Modica is a former head of the Conference of Italian University Rectors (an association of university chiefs), and now a member of Prodi's union. He says, for example, that the document's proposal for an independent authority to evaluate research done in all publicly funded research institutes and universities is now a central concept in the proposed government programme. "It would eliminate the unfair and inefficient elements in the Italian academic system," he says.

The scientists summoned by Prodi, none of whom is affiliated to a political party, argue that the problems have been there for decades, but have worsened in the past four years of Silvio

Berlusconi's rule. The government has reduced Italy's scarce science funds for basic research, they say, and oriented the sector to applied research. Berlusconi's centre-right coalition has not issued formal statements about science in the run up to the elections next month, but sources close to Berlusconi indicate a continuation of this philosophy.

Uphill struggle

"The situation wasn't good before, but the Berlusconi government made it much worse," says Giorgio Parisi, a theoretical physicist at the University of Rome, 'La Sapienza'. "Perhaps the worst thing that has happened has been damage to the research agencies," adds Massimo Inguscio, an atomic physicist at the University of Florence. He says a series of much needed but failed reorganization attempts over the past decade have left agencies like the National Research Council (CNR) and the National Institute for the Physics of Matter with unclear sets of rules, imposing a culture of extreme uncertainty, "which is not supportive of free research".

Others have openly criticized Berlusconi's set of politically appointed agency heads as scientific lightweights, who lack the charisma needed to defend the Italian tradition of research (see 'Careless with the truth').

Curiously, the problems of Italian scientists have not so far translated into an equivalent dearth of high-quality research. Given the difficult conditions in which many have to work, the output in fundamental research is relatively impressive; in number of publications and publication impact, Italy scores seventh out of the world's 140 highest-performing countries. Some point out, however, that the latest assessments include data that are already five years old, so may not reflect the harder times that have come recently.

As Italy spends only half the European Union average on research and development, the issue of money is on everyone's mind. But most agree that more money alone is not the answer. Fil-

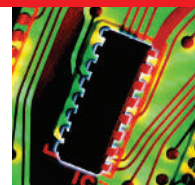


Until 1998, university jobs were announced centrally in Rome.

S. ALLEN TRAVEL PHOTOGRAPHY/ALAMY

ippo Andreatta, a political scientist at the University of Bologna and one of Prodi's chief advisers, says that although more money is foreseen — that indeed was the point of Prodi's thought experiment — funds will be limited as Italy is struggling to stay out of recession. More important, he adds, a centre-left government would "turn upside down how we give money".

In effect this means making new rules to ensure fair and effective competition for research jobs and funds. This may not be music to the ears of Italian scientists, who



AMERICAN PHYSICAL SOCIETY MEETING

Visit our website for blog reports, editors' picks of notable sessions and more.
www.nature.com/news



Arch rivals: Romano Prodi (right) is promising to rescue Italian science in a bid to oust Silvio Berlusconi.

despair of what they call *La riforma continua*, referring to the reforms to research organizations and recruitment procedures that began with Prodi's first term of office in 1996, and that have not stopped since.

The CNR, for example, which runs 100 research institutes, has had three major but ineffective shake-ups in the past decade. And, hampered by the confusion, universities are still trying to achieve the goals of a 1990 law that gave them autonomy to run their own budgets. Continuing uncertainty over rules has meant that few new research programmes have been launched.

But Rossella Palomba, a social demographer at the CNR Institute for Population and Social Policies in Rome, and a member of Prodi's group of scientists, says that "some restyling is unfortunately necessary, at least to ensure that academic recruitment allows us to hire the best people".

For those used to seeing academic positions advertised publicly as soon as they become available, the system of Italian recruitment

competitions, or *concorsi*, seems incomprehensible. Until 1998, all jobs at universities and research institutes were organized centrally in Rome; every year or two, there would be a single mass announcement, and phone systems would become jammed with personal lobbying. Since then, a series of changes have attempted to leave recruitment decisions to the institutions involved, while trying to discourage the tendency to recruit locally.

Chaotic inclination

But because many universities continued to fail to recruit outsiders, research and education minister Letizia Moratti reintroduced a type of centralized *concorsi* system, in a decree that was forced into law last November.

This recentralization would be reversed by a centre-left coalition, says Modica. "We need to create an environment where universities both have the right to choose their own candidates and feel compelled to choose the best." He favours "a controlled ten-year transition linked to evaluation, to allow universities to adapt to a new culture where they are penalized if they recruit poorly".

The Olive Tree coalition also foresees two other policy changes. One would tone down the emphasis on research with immediate applications, introduced by the Berlusconi government. This industry-friendly philosophy included reorienting the CNR's mission from fundamental to applied research. Despite its unpopularity among scientists, Fabio Pistella, appointed by the government as the CNR president in July 2004, says the focus on applications must continue. "Italian industry invests little in research and the CNR's mission is now to fill this gap — scientific papers are not the only measure of success of a good research organization."

The other shift in policy would reintroduce sources of research grants that have all but dried up, as well as more conventional mechanisms for distributing them. Even the CNR has virtually no money for actual research projects; most of its budget is eaten up by running costs.

The CNR researchers can team up with university professors who are eligible to apply for scarce university project grants distributed by the research ministry. But the ministry has also come under attack for inefficient management of the grants. One international research programme rejected 60% of its grant applications as invalid — including many from very experienced researchers.

Cell biologist Jacopo Meldolesi from the Vita Salute San Raffaele University in Milan,

CARELESS WITH THE TRUTH

Several high-ups in Italy's research organizations have come under fire for making dubious boasts.

● **The CNR president Fabio Pistella** claims to have 150 scientific publications on his CV; this was submitted to the Italian parliament in support of his presidential nomination in 2004. But *Le Scienze* reported in January 2006 that ISI cites only three publications by him. Pistella told *Nature* that some of his publications are old and in Italian, "and the roles of the CNR president in any case require management skills".

● **Claudio Regis, the vice-commissioner of ENEA, the Italian environmental agency**, uses the formal title 'Engineer'. But on 2 August 2005 the *Corriere della sera* newspaper reported that he is not listed as qualified with the Italian Guild of Engineers. Regis did not respond to requests for clarification.

● In the Bologna newspaper *Il resto del Carlino* on 3 January 2006, **Sergio Vetrella, the president of ASI, the Italian Space Agency**, was reported to have claimed that Italy will build a telescope on the moon. But the agency has no such plan. Vetrella told *Nature* he was misquoted, but the journalist argues that he approved the text. On 28 February, Vetrella announced a call from ASI's website for ideas for moon-related plans.

who is president of the Italian Federation of Life Sciences, had his application rejected before review for not meeting all criteria. This was despite him getting reassuring answers to his phone and e-mail enquiries. "Obviously the programme was not explained properly to applicants," says Meldolesi, who was also a member of Prodi's group of scientists.

"The situation wasn't good before, but the Berlusconi government made it much worse."

One concern that has found no consensus in Prodi's coalition is the issue of embryonic stemcell research. The research community would like to see more liberal rules, as the authors of the December document make clear. But Catholic and conservative coalition partners oppose loosening the embryo-protection laws, which are among the strictest in Europe.

Some things will have to wait until after the April elections, acknowledges Giovanni Bignami, an astrophysicist at the University of Pavia, who chaired Prodi's group of scientists. But he says: "We were happy to have been consulted by politicians. That hasn't happened in a while in Italy."

Biologists appeal to Senate to safeguard species act

More than 5,700 US biologists have signed a letter, delivered to the Senate on 8 March, urging senators to uphold the scientific principles of the Endangered Species Act. Bills to reform the act are currently being assessed by the Senate.

The act deals with the conservation of threatened and endangered plants and animals and their habitats. The House of Representatives has passed a reform bill, sponsored by Richard Pombo (Republican, California), which many biologists say would weaken the act's protective powers (see *Nature* 438, 140–141; 2005). That bill has now moved to the Senate, where a second, related bill is also being considered.

The Union of Concerned Scientists, which helped coordinate the letter, criticizes the legislation for putting scientific decisions in the hands of political appointees, changing the definition of a species' 'critical habitat', and excluding the use of common scientific techniques when making decisions that affect species.

Tap the power of the people to get clean water, says UN

The world has plenty of fresh water, but people still do not have equal access to it, says a report from the United Nations.

Released on 9 March, the study builds on the first *World Water Development Report* in 2003, which identified 11 areas where the world needs to improve water management. The latest edition concludes that bureaucratic restrictions are limiting people's access to water. Citizens, the report argues, need to be allowed to take responsibility and organize themselves to ensure access to sufficient clean water.

The UN's millennium development goals



Hard to swallow: many people around the world still do not have access to clean, fresh water.

Yeti discovered near South Pacific island

A crustacean dubbed the Yeti crab (*Kiwa hirsuta*) has stumbled into the spotlight. This furry white crab (pictured) hails from a remote spot more than 1,000 kilometres south of Easter Island in the Pacific Ocean.

Scientists led by the Monterey Bay Aquarium Research Institute in California found the 15-centimetre-long creature at the periphery of a newly discovered hydrothermal vent some 2,200 metres below sea level. Studies revealed it is distinct enough from its known relatives to form the first member of a



A. FIFIS/IFREMER

new family: the Kiwaidae. In Polynesian mythology, Kiwa is the goddess of shellfish.

The blind crab's 'hairs', which feel like toothbrush bristles, are covered by filamentous bacteria, but their function is not yet known.

include halving the number of people who do not have access to safe drinking water or basic sanitation by 2015.

Pentagon permits inquiry into classified data

The US Department of Defense is to investigate allegations of data tampering at the Lincoln Laboratory, part of the Massachusetts Institute of Technology (MIT).

After a series of missile-defence tests in 1997 and 1998, data were allegedly altered to hide the failure of a sensor. MIT stopped its investigation after the Pentagon refused to turn over what it said were classified data (see *Nature* 432, 789; 2004).

Now, MIT says that the Pentagon has agreed to conduct an investigation, to be led by Kenneth Krieg, the under-secretary of defence for acquisition, technology and logistics.

"To me it's astonishing," says MIT professor Ted Postol, an outspoken critic of missile defence. "They've turned over management of MIT's academic integrity to the defence department."

But Claude Canizares, an associate provost at the university, says that the classified nature of the materials left them little choice. "This is not optional; it's part of federal regulations," he says.

Britain pursues plan to scale back climate centre

Plans to close four UK environmental research sites, which scientists say are needed to help understand the threat of climate change, have been approved with only minor modifications.

The Natural Environment Research

Council confirmed on 8 March that four of the Centre for Ecology and Hydrology's eight sites will close as planned. Two hundred positions had been expected to go, but the council says up to 40 of these could be saved by its decision to increase the centre's target for bringing in external research funds from £11 million (US\$19 million) to £12.4 million per year.

The council also pledged to maintain operation of the centre's long-term data sets. Some of its records on biodiversity and water quality go back decades, and many researchers had feared that other funders, such as universities, would be reluctant to take them over if the research council withdrew funding (see *Nature* 439, 770–771; 2006).

US drug agency looks set to get permanent leader

The acting commissioner of the US Food and Drug Administration (FDA) is likely to be nominated for the permanent position imminently, according to a knowledgeable source.

Andrew von Eschenbach, a urologic surgeon, took over the agency on a temporary basis after the sudden resignation of Lester Crawford last September (see *Nature* 437, 606; 2005). To become permanent commissioner, he would need to be nominated by President George W. Bush and then confirmed by the Senate.

Von Eschenbach is still nominally chief of the National Cancer Institute — he was appointed director in 2002. But he has surrendered the institute's day-to-day management to John Niederhuber, one of the deputy directors.

The FDA has been without a permanent head for more than three of the past five years.

LET THE GAMES BEGIN

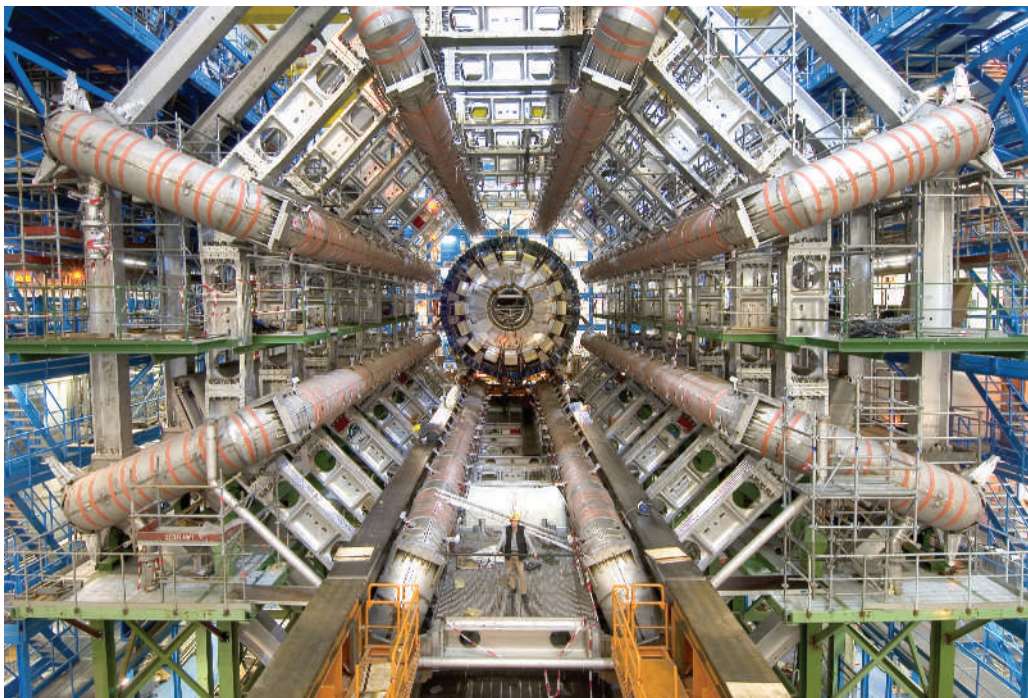
A series of mental challenges is helping physicists to prepare for the strange data they may get when the next particle accelerator goes live. **Jenny Hogan** joins the work-out.

Normally, the trick to learning something at a scientific meeting is to listen to the key lectures. But one afternoon last month, in a conference room at CERN, the European particle-physics lab near Geneva, physicist Matt Strassler managed to convince several researchers that they might learn more if they left the lecture room. He wanted them to avoid hearing the solution to a puzzle they had been working on for months.

Welcome to the strange world of the Large Hadron Collider (LHC) Olympics, a workshop held at CERN in which teams of theorists studied fake data in order to explore unproven theories.

Strassler, a theorist from the University of Washington in Seattle, was one of the organizers of the event, which brought together more than 50 theoretical physicists from across Europe and the United States. These 'olympians' have devoted their careers to building mathematical models of the Universe and matter. The LHC Olympics was designed to put their ideas to the test; their challenge was to prepare for the real data expected to emerge from one of the biggest experiments in physics. Strassler and his colleagues had accepted that challenge—and they weren't going to give up on their first attempt.

That experiment is the LHC, a machine under construction in the tunnels beneath CERN. When it is switched on in 2007, the collider will be the world's most powerful particle accelerator; it is expected to churn out data that will test some of the most cherished theories of physics. But if something turns up in the LHC that no one has predicted, theorists



Monster machine: huge detectors will collect and characterize the fall out from proton collisions.

CERN

will have to work backwards from a set of observations to try to find a model that fits.

"Everyone in this room wants to be prepared so that when the first slide of the first data from the LHC is presented, we will understand what we are seeing," says Jesse Thaler, a physics PhD student from Harvard University in Cambridge, Massachusetts.

Studying the debris from collisions in particle accelerators has already helped physicists piece together many details of the subatomic

world. But the LHC is pushing particle physics to a new frontier.

When high-speed protons smash head-on in the LHC, obliterating each other, seven times more energy will be released than from similar collisions in today's most powerful machine, the Tevatron at Fermilab in Batavia, Illinois. This raises the possibility that in the explosion of particles that sprays out, some new, exotic speck of matter might be found.

Full speed ahead

The experimentalists building the mammoth detectors that collect and characterize the debris will comb through the data, searching for new particles. They have many ideas of what they might find at these energies, but it is possible that there may also be some complete surprises. "To go from seeing particles to understanding theories is a big step. This is where the LHC olympians will come in," says Albert de Roeck, physics analysis coordinator for the LHC detector known as the Compact Muon Solenoid (CMS).

Physicists have not had to contend with unexpected physics since the 'standard model', which divides matter into families of particles, was hammered out in the 1970s.

"The 1960s and 1970s were the golden years for particle physics," says de Roeck. "The energy that was there then hasn't really been there



Bright sparks: Nima Arkani-Hamed (second from left) helped organize a collider workshop at CERN.

CERN

since. The hope of the whole particle-physics world is that the LHC will bring that era back.”

The LHC Olympics was also designed to attract new people into the fray. Herman Verlinde, a string theorist from Princeton University, New Jersey, is one of the newcomers who helped organize the meeting. “String theorists, like any physicists, are attracted to exciting areas of physics,” he explains.

String theory is one contender for a deeper theory of nature, which goes beyond standard-model physics. Despite its unqualified success, the standard model contains at least 19 arbitrary parameters such as particles’ masses that have to be measured from experiments. Physicists seek some greater underlying theory to explain these numbers, and have dreamed up numerous formulations that might provide it.

Going for gold

Some of these theories predict that the standard model will be supplemented by families of heavier particles, which the LHC may find. And the LHC could identify dark matter, uncover evidence for extra dimensions, or create miniature black holes. At the very least, physicists hope it will find the Higgs boson — a hypothetical particle, predicted to exist by the standard model, which is thought to endow other particles with mass. Or it might find nothing at all.

“I can’t emphasize enough how tense and exciting this time is,” says Nima Arkani-Hamed, a theorist from Harvard University, and one of the driving forces behind the LHC Olympics.

Last time CERN made a big discovery, Arkani-Hamed explains, the theorists had known exactly what the experimentalists should look for. “This time,” he says, “the theorists are saying ‘maybe this will happen, maybe that will happen.’”

Arkani-Hamed first became worried in 2004 that many theorists were not prepared for the flood of data that will soon issue from CERN. In his original vision for the LHC Olympics, experimentalists would create a simulated data set that obeyed some hidden physics from beyond the standard model, analyse it as they would the real data, and then present their plots to an audience of theorists.

“The idea was that we would have a massive data challenge and then a mini conference,” says Maria Spiropulu, who will work on the CMS data analysis at CERN. At the conference, the theorists would be challenged to explain the experimentalists’ plot, “Then they would say, ‘Eureka! This is intersecting D-brane string-inspired supersymmetry.’”

But right now, as the start date for the LHC looms, the experimental teams are too busy to get involved with games. “We’re working 16-hour days as it is,” says Spiropulu. “But these people don’t give up. The theorists said, ‘If you give us a hint, we’ll do it ourselves.’”

So, a few months before the February workshop, several theorists put together three simulated data sets. These ‘black boxes’ were then posted on a website for other meeting participants to download. Could the conference-goers work out, from the plots and tables, what physics the data described?

As you might expect, says Spiropulu, the theorists’ data analysis was very basic. “But for them, this was training so that they understand the language that we are using.” With real LHC data, one of the big problems will be distilling the bumps that signify new particles from the ‘background’ events produced by standard-model physics. No background was included in

the theorists’ black boxes.

Ian Hinchliffe at the Lawrence Berkeley National Laboratory in California, a physicist on the ATLAS detector team, worries about this omission. “It might lead some theorists to wrong conclusions about what is possible,” says Hinchliffe, who ran a data challenge for the people working on ATLAS, which included the necessary background events. That experience, says Hinchliffe, taught them how hard it might be to recognize new events.

Runners up

But the LHC Olympics is not a bad starting point, says Strassler. He describes the black-box exercise as “a sort of flight simulator for someone who wants to understand clearly what the pilots are doing, rather than something for the pilot”. And for the next meeting, planned for August, the theorists hope to develop at least one black box with background.

First to present his black-box solution this time was Verlinde. “Clearly we are not going to impress you with our skills

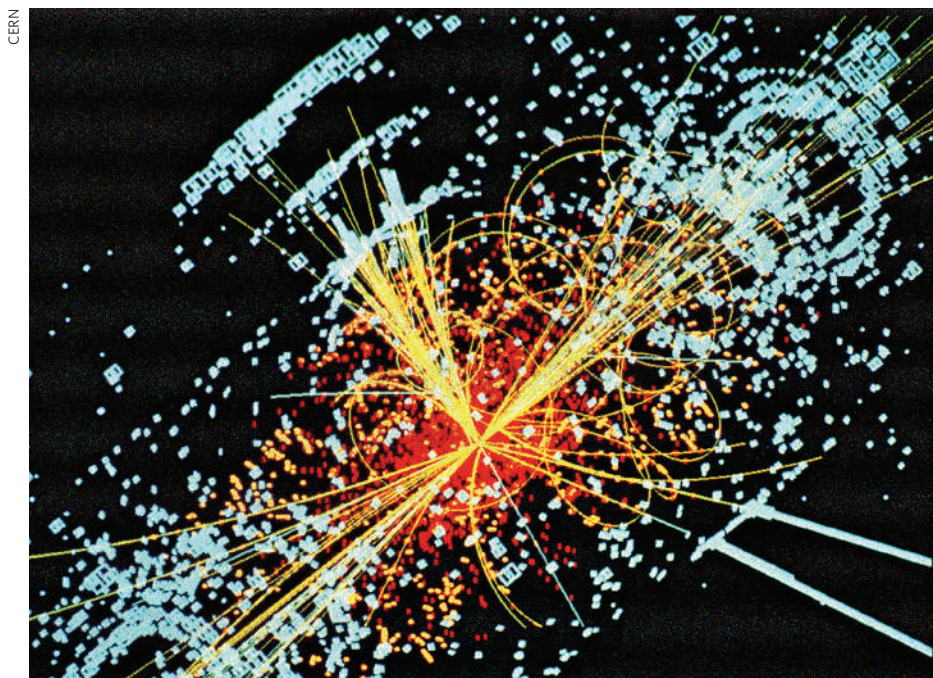
in analysing data,” he began. “You can see how far one can get, starting basically from zero, with minimal students and minimum sleep over the past few days.”

Another presentation came from Harvard University students, who had tested 2,500 models, using an algorithm to optimize their first guess of what particles in the black box might be. “Cheaters,” someone called out when the team thanked the Harvard-Smithsonian Center for Astrophysics for allowing them to use its powerful computers; the heckler’s own talk was titled “What you can do with a black box, an undergraduate and two weeks”.

Both Verlinde and his co-workers, and the Harvard students had figured out that the box they had studied was based on a version of supersymmetry — among the leading theories for a physics beyond the standard model. And they identified the masses of most of the particles that this theory predicts — from squarks, to winos and zinos. There were some differences between the two groups, but their answers were almost the same. “We’re quite proud of our guess,” says Princeton University theorist Leonardo Rastelli who collaborated with Verlinde.

But Strassler hadn’t figured out what was going in the black box he had tackled. That’s why, when its solution was to be announced, he persuaded several others who had been working on it to leave the room with him. He wanted to keep going until he cracked it. “If the black box hasn’t been decoded, then the learning process isn’t over.” And of course, when the real results come in, there won’t be anyone who can reveal the answer. It will have to be obtained through hard work. ■

Jenny Hogan is a reporter based in *Nature*’s London office.



High-impact: when protons collide in the LHC, their energy will be converted into a spray of particles.



STACKING THE DECK

Studies of medical literature are confirming what many suspected — reporters of clinical trials do not always play straight. **Jim Giles** talks to those pushing for a fairer deal.

They answer only the questions they want to answer. They ignore evidence that does not fit with their story. They set up and knock down straw men. Levelled at politicians, such accusations would come as no surprise. But what if the target were the researchers who test drugs? And what if the allegations came not from the tabloid press, but from studies published in prestigious medical journals?

The slurs may sound over the top, but each is based on hard data. Since 1990, a group of researchers has met every four years to lay bare the biases that permeate clinical research. The results make for uncomfortable reading. Although outright deception is rare, there is now ample evidence to show that our view of drugs' effectiveness is being subtly distorted. And the motivation, say the researchers, is financial gain and personal ambition.

"Patients volunteer for trials, but finances and career motives decide what gets published," says Peter Gøtzsche, an expert in clinical trials and director of the Nordic Cochrane Centre in Copenhagen. "This is ethically indefensible. Change is not easy, but we must get there."

It is a dramatic conclusion to come from a field of study with no proper name, staffed by

part-time volunteers. Most are journal editors, medical statisticians or public-health experts, united by fears for the integrity of clinical trials. For the devotees of 'journalology' or 'research into research', the literature on clinical trials is their raw data and patterns of bias are their results.

Some of these researchers are using their findings to change medical journals and make it harder for authors to misrepresent results. Others are working on what could become the biggest reform of clinical-trial reporting for decades: the creation of a comprehensive international registry of all clinical trials. It is a powerful idea, which could one day make all trial information public. It is also an idea that has pitched pharmaceutical companies against advocates for reform, in a tussle over whether transparency or commercial confidentiality best serves medical science.

Just say no

One of the biggest problems with clinical-trial reporting, the suppression of negative results, shows the importance of such debates. Because clinical researchers are not obliged to publish their findings, ambiguous or negative results can languish in filing cabinets, resulting

in what Christine Laine, an editor at the *Annals of Internal Medicine* in Philadelphia, Pennsylvania, calls "phantom papers". If that happens, the journal record will give an over-optimistic impression of the treatments studied, with consequences for peer reviewers, government regulators and patients.

One alleged example hit the headlines in 2004. At that time, the antidepressant Paxil (paroxetine), made by London-based drug giant GlaxoSmithKline, was a popular treatment for adolescents in the United States. But doctors have now been warned off prescribing Paxil to youngsters, after evidence emerged that it increases the risk of suicidal behaviour. It was claimed in a court case brought in the United States that GlaxoSmithKline had suppressed data showing this since 1998. Rick Koenig, a spokesman for GlaxoSmithKline, says the company thought the charges unfounded, but agreed to pay \$2.5 million to avoid the costs and time of litigation.

Phantom papers can be tracked down through trial protocols — the document describing how a trial will be run and what outcomes will be measured — which have to be registered with local ethics committees. By matching papers with protocols, several groups

C. DANKIN

have shown that many trials are completed but not published. And that, notes Laine, makes it impossible for journals and health agencies to assess potential drugs. "You never quite know if other data are out there that would influence your conclusions," she says.

Last year, for example, a French team showed that only 40% of trials registered with its country's ethics committees in 1984 had been published by 2002, despite more than twice as many having been completed¹. Crucially, papers with inconclusive results not only took longer to publish (see graph), they were less likely to see the light of day at all. Researchers in any field can sit on negative or inconclusive results. But critics say that clinical researchers carry a greater ethical burden, as their findings inform decisions about the licensing of drugs.

Don't believe the hype

Nor do the problems end when a trial hits an editor's desk. Results from a trial of the arthritis drug Celebrex (celecoxib) looked good when they were published in 2000, for example, but less so when physicians scrutinized the full data set. The original paper, which appeared in the *Journal of the American Medical Association (JAMA)*, dismissed fears that Celebrex could cause ulcers. But that was based on data collected over six months. When other physicians analysed a full year's worth — which the authors already had at the time of their *JAMA* submission — they claimed that Celebrex seemed to cause ulcers just as often as other treatments². The original study's authors say that the later data were too unreliable to be included, but acknowledge that they could have "avoided confusion" by explaining to editors why they had omitted them.

This case is not a one-off. During his PhD at the University of Oxford, UK, epidemiologist An-Wen Chan looked at the protocols for 122 trials registered with two Danish ethics committees in 1994–95. More than half of the outcomes that the protocol said would be measured were missing from the published paper, he found³. Asked about the missing outcomes,



Bitter pill? GlaxoSmithKline denies suppressing data about antidepressant Paxil's side effects.

most authors simply denied the data were ever recorded, despite evidence to the contrary. And when Chan looked at those missing data, he found that inconclusive results were significantly more likely to have been left out of the final publication.

Medical journals can, however, request such missing data. One idea, adopted by *The Lancet* three years ago, is to insist that authors send in a trial protocol when they submit results. That should help identify whether researchers are reporting all the information they gathered.

But even with all the data, journal editors face another challenge: hype. Researchers and sponsors tend to be interested in things that work rather than those that do not, so authors may subconsciously tweak results and talk up conclusions. "Researchers are so worried about getting papers rejected that they put a lot of spin on results to make them seem as exciting as possible," says Doug Altman, a medical statistician who supervised Chan at Oxford.

The hype shows up in a paper's conclusions. In 2003, epidemiologist Bodil Als-Nielsen and her colleagues at the University of Copenhagen looked at factors that might influence

researchers' conclusions about a drug's efficacy or safety⁴. Their analysis of 370 trials showed that the strongest predictor of the authors' conclusions was not the nature of the data, but the type of sponsor.

Trials funded by for-profit organizations were significantly more likely to reach a favourable verdict than those sponsored by charities or governments. Critically, the association was not explained by the papers having more positive results. In a study under review, Gøtzsche and his colleagues show that industry-funded meta-analyses — studies that combine results from several clinical trials of a drug — are similarly prone to draw positive conclusions that are not supported by the data (see graph).

For many clinical-trials experts, these funding biases explain all the others. For each act, be it the suppression of results or the omission of outcomes, there is a financial motive for the company whose drug is being tested. In many cases, the company funding the study also employs one or more of the authors. Given the combination of motive and opportunity, many see drug-company influence as an inevitably distorting factor.

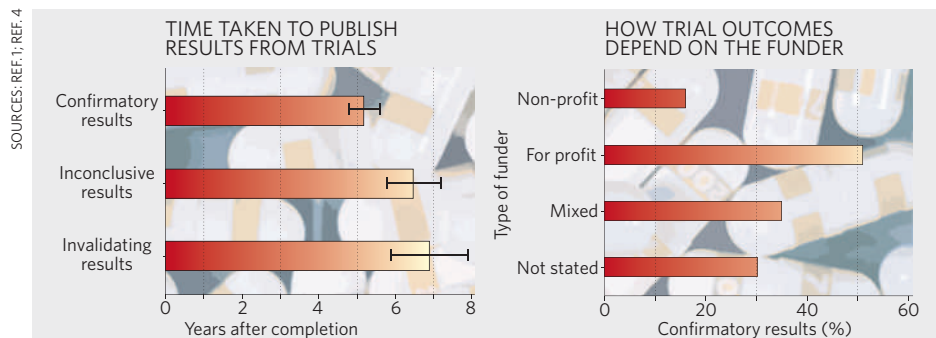
"When we see an industry article we get our antennae up," says Steven Goodman, a medical statistician at Johns Hopkins University in Baltimore and an editor at the *Annals of Internal Medicine*. "It's not that we assume the research is done badly. But we have to assume that the company has done all it can to make its product look as good as possible."

Editorial control

At *JAMA*, editors began insisting last year that all research sponsored by for-profit organizations undergo independent statistical analysis before acceptance. Cathy DeAngelis, the journal's editor-in-chief, says *JAMA* had asked authors to do this for years, but began requiring it after editors started seeing papers that they thought dishonest. "People said that for-profit companies would stop sending us trials," she notes. "Well, guess what? If you look at what we're publishing you'll see that that's not true."

Still, Goodman and others caution against blaming everything on industry. Government-sponsored trials also tend to report positive outcomes⁵, although the effect is weaker than with industry studies. And a publication in a big journal can boost authors' careers as well as company coffers.

Others add that journals must also share the blame. Good peer reviewers and hands-on editors should, for example, weed out hype. But according to Richard Smith, a former editor of the *BMJ* (which was the *British Medical*



JACOB RIIS



"Patients volunteer for trials, but finances and career motives decide what gets published."

— Peter Gøtzsche

Journal) and now head of European operations for the US insurer UnitedHealthcare, editors may be biased towards positive results. In an article published last May, titled "Medical journals are an extension of the marketing arm of pharmaceutical companies", he pointed out that reprints of papers reporting positive results can generate millions of dollars, and that this might influence editorial decisions⁶.

There is certainly evidence that drug companies attempt to use reprint income as a lever on journals. *The Lancet's* Richard Horton has said that authors sometimes contact him to say that sponsors are likely to buy large numbers of reprints if their study is published. Horton and other editors at top journals say they rebuff such threats, but some less well staffed journals lack policies for separating commercial and editorial decisions, suggesting that reprint income at least has the potential to distort decisions.

Merrill Goozner, who tracks pharmaceutical issues at the Center for Science in the Public Interest, a lobby group in Washington DC, agrees. "It's a financial conflict of interest, plain and simple," he says.

Full disclosure

Such accusations make medical editors angry. They deny that commercial pressures influence peer review, adding that journals have introduced several measures that have helped to clean up clinical-trial reporting.

One of the first initiatives, introduced in 1996 and revised in 2001, is the statement on Consolidated Standards of Reporting Trials (CONSORT). A set of guidelines on how to report a clinical trial, the statement is designed to ensure that authors present results transparently. It seems to be helping. Since top journals began insisting that authors follow the guidelines, researchers' descriptions of their methods, for example how they place subjects in treatment or control groups, have become more accurate⁷. Such information aids reviewers' decisions.

Journals have also endorsed trial registries.

By registering all trials when they begin, researchers will find it harder to suppress outcomes, editors believe. Several registries already exist, including one run by the US government. The World Health Organization (WHO) is working on an online portal that would bind these databases into a single source. And in 2004, the International Committee of Medical Journal Editors announced its members would not publish the results of trials that had not been placed in a public registry.

Clinical-trials experts welcomed the move, but the industry response was patchy. Last June, the committee was forced to issue another statement after finding that some sponsors were being deliberately vague and entering terms such as 'investigational drug' in the field for the drug name. A follow-up study found the quality of information had improved considerably by last October⁸.

Despite these successes, advocates of reform say bigger fights lie ahead. The experts working on the WHO registry want a list of mandatory entries for trial data, including the primary outcome. For drug companies this is a step too far, akin to asking an inventor to publish the description of an invention before it is patented. Instead, the companies propose depositing such information in a locked database, to be released when the drug obtains marketing approval. Going public too early,

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

On the record: journals have begun demanding more transparent registries of clinical trials.

"We have to assume that the company has done all it can to make its product look as good as possible."

— Steven Goodman



S. GOODMAN

they say, would deter companies from taking risks on potential treatments and slow down the generation of new drugs.

Yet for critics such as Smith, even the WHO portal does not go far enough. Along with other registry advocates, he would like to see all clinical results, not just protocols and outcomes, published in public databases.

This proposal would seem to tackle problems with reporting data and bias. But it would not be simple. Aside from the commercial concerns of the drugs industry, the creation of a results database could lead to patients pressurizing doctors for access to experimental medicines. Health insurers and hospitals might also change the drugs they use after seeing the raw results, rather than waiting for peer-reviewed papers.

The debate is in its infancy. Yet clinical-trials experts are more optimistic than they have been at many points in the past 15 years. Chan, now working for the WHO registry team in Geneva, says the first round of consultations with stakeholders should produce a policy statement in April. He and others add that although industry will probably continue to resist, the public attention generated by recent scandals, and the wealth of data available on the problems, mean that time is ripe for change.

"I'm not relying on hope," says Gøtzsche. "But the results of all trials should be made public, not only those that the sponsor cares to tell the world about. This is incredibly important." ■

Jim Giles is a senior reporter for Nature in London.

- Decullier, E., Lhéritier, V. & Chapuis, F. *Br. Med. J.* **331**, 19–22 (2005).
- Hrachovec, J. B. & Mora, M. *J. Am. Med. Assoc.* **286**, 2398 (2001).
- Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C. & Altman, D. G. *J. Am. Med. Assoc.* **291**, 2457–2465 (2004).
- Als-Nielsen, B., Chen, W., Gluud, C. & Kjaergard, L. L. *J. Am. Med. Assoc.* **290**, 921–928 (2003).
- Dickersin, K. & Min, Y. I. *Online J. Curr. Clin. Trials* **2**, 50 (1993).
- Smith, R. *PLoS Med.* **2**, e138 (2005).
- Moher, D., Jones, A. & Lepage, L. *J. Am. Med. Assoc.* **285**, 1992–1995 (2001).
- Zarin, D. A., Tse, T. & Ide, N. C. *N. Engl. J. Med.* **353**, 2779–2787 (2005).

R. FARIS/CORBIS

Q&A

MAKE A STRONG BOND

Alison Abbott talks to the man who wants theoretical chemistry to ease political strife in the Middle East.

Can scientific ties counter political tensions? Roald Hoffmann hopes so. This Nobel laureate from Cornell University, New York, has organized a series of three chemistry workshops for young scientists from the Middle East, with the aim of forging trust and friendship between participants. The idea stemmed from a small international chemistry meeting in Malta in 2003, which was attended by many chemists from Arab countries.

Hoffmann taught the first of his workshops, called 'Chemistry Bonds', in Jordan's ancient city of Petra in January. The 13 attendees hailed from Jordan, Palestine, Saudi Arabia and Syria, as well as from Israel and Iran.

Political relations in the Middle East are very tense. Can academic classes really help?

It is unrealistic to think a scientific gathering will solve the problems of the region. But perhaps many small actions that build trust and connections between people, especially young people, will help. This is my small contribution to peace, but I'm not romantic or do-gooding about it. The group I brought together could be future leaders of science.

So how could chemistry possibly build trust?

Pere Alemany, my co-teacher from the University of Barcelona, and I worked those young people hard. So there was a certain shared suffering, and victory over that, on their part. Camaraderie builds in such intense situations. But we also did things that bring people close. One day we all cooked a Jordanian meal together; on another, we all trekked through the pink sandstone monuments of Petra. I asked each participant to bring a favourite piece of music. For me, music is the next best thing to science for bringing people together. At 10 p.m. each night we listened to the tapes with drinks: wine, soft drinks. Alcohol wasn't an issue. Some drank it, some did not. I wasn't going to go without wine!

Was religion an issue?

No, not really, but it surprised me how little the participants knew about each others' customs. Most were from Muslim backgrounds, a few were Jewish and there was at least one Christian. Some were secular, some religious. One of the Jewish students ate kosher, which led to illuminating discussion of dietary rites.



Summit meeting: Roald Hoffmann, in the red scarf, brings together young chemists from the Middle East.

Was gender an issue?

Only in that it showed the ignorance we have in the West. Half of the applicants were women.

Did the participants get on with each other?

Yes, although at the beginning there was much shyness. I learned in Malta how little communication there actually is between scientists from the various Arab countries. But I was taken aback when one Israeli confessed she had never had a social conversation with an Arab, even a Palestinian, before she came to the Petra workshop. Several of the participants said something changed in them that week.

Do you think they will stay in contact with each other?

I hope so. We will help them, but I think it will come naturally. The bonds that formed are strong; I feel that. I was touched that one of the Jordanians called the conference coordinator, Vanessa Buisson, to find out if the Israelis had got home alright despite a bombing in Tel Aviv on the day of departure. He said: "They grew part of me that week."

Did the participants risk attack or denunciation by taking part in the workshop?

Part of me says there was no risk; part says, be realistic. Let's be frank: the risks were personal, perhaps greatest to the three Israelis and the American organizers. But in a different way

there were risks to students from the Arab countries. They spent time with Israelis; some people back home don't like that.

How much chemistry did they learn?

A lot. Everything that Pere and I had to teach them from a lifetime of molecular orbital lore. It will serve them well. Only one or two will become theoretical chemists. The rest will use their learning in whatever they do. I was touched when at the end of the course, one student thanked me for "making chemistry come alive for me again".

Where will the next two workshops take place?

In the next 18 months, Harvard's George Whitesides will lead one in Egypt on nanochemistry and Harry Gray from the California Institute of Technology will teach another on bioinorganic chemistry in Qatar. We could teach more. My dream is to convince a Saudi prince to host a workshop for everyone in the region — for his young people, Israelis and others.

'Chemistry bonds' — is a metaphor intended?

Atoms bond because they don't have a choice. That's not bad; good chemistry comes from that bonding. But people do have a choice — and we gave some of them the opportunity to do so.

Alison Abbott is Nature's senior European correspondent.

BUSINESS

Make or break time in Vioxx drama

Cases involving long-term users of Vioxx will, as **Meredith Wadman** reports, determine the true cost to Merck and the drug industry of the painkiller's withdrawal.

In a trial that opened on 6 March in Atlantic City, New Jersey, lawyers led by a charismatic Texan are trying to convince a jury that Merck's blockbuster painkiller caused heart attacks in two allegedly long-term users.

It is the first time that the company has confronted plaintiffs who have taken the drug for more than 18 months — the period after which, according to the study that led Merck to pull the drug in September 2004, Vioxx boosts the risk of heart attacks and strokes.

That makes the stakes in the current trial arguably the highest yet. "A win for Merck may deflate the entire plaintiff case," says Anthony Butler, an analyst who is following the litigation for the investment bank Lehman Brothers in New York.

The withdrawal of Vioxx, a leading arthritis treatment that earned Merck \$2.5 billion in 2003, has so far led to nearly 10,000 lawsuits against the company. In the first of those, last summer, a jury in the Texas town of Angleton awarded \$253 million to the widow of a man who had a fatal heart attack after taking Vioxx for eight months (see *Nature* **436**, 1070; 2005). The award is likely to be cut to \$26.1 million under a Texas law capping punitive damages.

But Merck bounced back, winning the second and third cases. The second, in which a jury found that Vioxx was not responsible for the heart attack of a 60-year-old postal worker who took the drug intermittently for two months, was heard in the same New Jersey state court as the current cases. In the third, a jury in a federal court in New Orleans last month took just three hours to clear Merck of responsibility for the death of a Florida man who suffered a fatal heart attack after taking Vioxx for less than a month.

These verdicts have helped Merck to recover a half of the steep drop in its share price that accompanied Vioxx's withdrawal (see graph). In combination with the performance of other parts of its business, some observers can see light at the end of the tunnel for the beleaguered drug maker. "At the end of the day you need to separate Vioxx from the fundamentals of the company," says Butler. "It's about the ability to pay dividends, the ability to fund your R&D and to fund your new launches. And Merck seems to be funding its programmes regardless of the outcome of this case in Atlantic City or any in the future."



J.F. MORENO/AP

Away match: Mark Lanier hopes to repeat his Texan victory over Merck in the company's home state.

Others are less sanguine. Late last month, the Arlington, Virginia, stock-research company Friedman Billings Ramsey lowered its rating of Merck's stock, citing Vioxx litigation risk among its reasons.

Moving targets

Published estimates of the company's probable total liability vary wildly, from \$4 billion to \$50 billion. This uncertainty is reflected in Merck's share price, says Richard Evans, an analyst who follows Merck for Sanford Bernstein, a stock-research firm in New York. "Merck's ultimate liability remains very hard to pin down," he says.

In the latest trial, sometime Baptist preacher Mark Lanier, the folksy Houston lawyer who won the Texas award last year, is representing Thomas Cona, the owner of a vascular

ultrasound company, who was 57 when he had a heart attack after taking Vioxx for, he says, 22 months.

But Lanier is in Merck's home state this time. And both Cona and the other plaintiff, retired insurance agent John McDarby, survived their heart attacks, potentially limiting jury sympathy. McDarby was 75, and had been taking Vioxx for four years when he suffered a heart attack after hip surgery. Both men were former smokers with high blood pressure, high cholesterol and other cardiac risk factors.

Lanier acknowledges Cona's poor heart health, but argues that this is precisely why he should never have been prescribed Vioxx. "Of course they were at risk," he says. "Vioxx shoves you toward a heart attack. And it is those at risk who can least afford a shove."

Whether juries will buy that argument remains to be seen. The issues will broaden out in June, when an Alabama court hears the first case concerning Celebrex, a Pfizer painkiller from the same family as Vioxx and the only drug of its type still on the market.

In the meantime, two victories for Merck in Atlantic City could mark a turning point in the Vioxx saga. Even one loss, however, will boost the growth of a pool of plaintiffs that could ultimately number in the tens of thousands. ■



Wiki ware could harness the Internet for science

SIR — Your News story “Experts plan to reclaim the web for pop science” (*Nature* **439**, 516–517; 2006) describes a project called the Digital Universe, which aims to create, and provide links to, trustworthy peer-reviewed content on the Internet.

As suggested by critics in your News story, it seems wasteful to try to reproduce content that already exists in an open, accessible and improvable form. I would encourage those working on projects such as the Digital Universe to consider a strategy that truly leverages the power of the Internet. For example, the free and editable encyclopedia Wikipedia (en.wikipedia.org) contains innumerable articles that are scientifically accurate, although it obviously contains errors, omissions and some articles of low quality. But MediaWiki, the software upon which Wikipedia is based (see www.mediawiki.org), allows one to link to specific versions of articles. Thus, expert peer reviewers could analyse articles, improve them and provide links to the trusted version of that article.

A web portal that provided a ‘filtered’ version of the Internet, with links to the most trustworthy available article on a given subject, would be a boon to scientists and science enthusiasts alike.

Kevin Yager

Department of Chemistry, McGill University, Lab 406, Otto Maass Chemistry Building, 801 Sherbrooke Street West, Montréal, Québec H3A 2K6, Canada

Lyme vaccine demonized by advocacy groups

SIR — As a microbiologist who managed a federal programme on Lyme disease in the 1990s, I consider that any new clinical trials of a vaccine candidate based on the protein OspA, as mentioned in your News Feature “Uphill struggle” (*Nature* **439**, 524–525; 2006), should be confined to Europe, for three reasons.

First, Lyme disease is non-communicable, readily treatable with common antibiotics and geographically localized in the United States. Neurological cases — where treatment can be problematic — are more common in Europe and a new vaccine may reduce the costs and consequences of infection.

Second, European experience with the widely used tick-borne encephalitis virus (TBEV) vaccine may facilitate vaccine-trial recruitment and greater public acceptance of a new Lyme vaccine.

Third, Europe is a less litigious environment and is largely free of organized Lyme-patient advocacy groups. In the United States,

activists have turned Lyme disease into everyone’s backyard bogeyman. They have demonized experts for their views on treatment and prevention, and hired lawyers to successfully argue the dangers of vaccine-induced autoimmunity (*Philadelphia Inquirer* **B03**, July 9 2003).

The activists are already using Internet discussion groups to warn against a new vaccine. One of them recently wrote “I would encourage all Lyme patients to consider writing letters, emphasizing the lack of demand for the last vaccine, and also the fact that any future vaccines can expect a lack of cooperation, protests, legal quagmires, etc.”

A careful, hysteria-free trial of the new OspA vaccine in Europe may help to undermine the opposition to it in the United States.

Edward McSweeney

1692 Barrister Court, Crofton, Maryland 21114, USA

Lyme vaccine: studies have raised genuine concerns

SIR — The lamentable hostilities between researchers into Lyme disease and the patients on whose behalf they purportedly labour will not be eased by *Nature*’s account of the issues surrounding GlaxoSmithKline’s decision to pull its vaccine, LYMERix, from the market (“Uphill struggle” *Nature* **439**, 524–525; 2006).

The News Feature portrays truth-seeking researchers on one side and angry patients and their advocates on the other, with the latter torpedoing a perfectly good vaccine with scare tactics and bad publicity. But a significant portion of the Lyme research community had strong reservations about the product. LYMERix’s initial approval by the vaccine advisory committee of the US Food and Drug Administration (FDA) was hedged with caveats that questioned its safety and even its usefulness (*J. Am. Med. Assoc.* **279**, 1937–1938; 1998). One case series in the peer-reviewed medical literature describes a group of six patients, with no history of neurologic problems, who developed neuropathy or cognitive impairment shortly after receiving the vaccine (*J. Peripher. Nerv. Syst.* **9**, 165–167; 2004).

And there were other factors that hindered its success in the marketplace. It was relatively expensive at \$50 per inoculation, and it required three doses, spaced over a period of one year, to achieve its full efficacy rate, which topped out at only 79%. Perhaps most disappointingly, study subjects lost their immunity relatively quickly, meaning that booster shots would be required every year or two — shots that the FDA’s advisory committee refused to sanction without further study (www.fda.gov/ohrms/dockets/

ac/98/transcript/3422t1.pdf). In addition, the vaccine was not approved for children under the age of 15.

Carl Brenner

Lamont-Doherty Earth Observatory of Columbia University, Route 9W, Palisades, New York 10964, USA

Reprocessing method could allay weapons fear

SIR — We believe that your worry about US plans to reprocess nuclear fuel (“Recycling the past” *Nature* **439**, 509–510; 2006) is misplaced. Since President Carter imposed a reprocessing ban in 1977, it has become clear that other nations’ decisions about building nuclear weapons do not depend on what the United States does with its spent fuel.

Furthermore, we consider your claim that “recycling involves separating components that can readily be used to build nuclear weapons” to be misleading on two counts. First, degraded plutonium in spent reactor fuel can only be used in an explosive device with considerable difficulty. Second, although current recycling processes produce pure plutonium that can be used for weapons, the US plan is to perfect a new method called UREX+, which would be configured so as never to separate weapons-quality plutonium.

UREX+ processing is the first step towards consuming excess plutonium in advanced, metal-fuelled fast reactors and reducing the rate at which reactor-grade plutonium is accumulating around the world. Moreover, fast reactors can extract more than 99% of the energy in mined uranium — over a hundred times better than the thermal reactors used today. The combination of recycling and fast reactors also reduces the time that waste needs to be isolated, from thousands of years to a few hundred.

There is still the associated problem that the Nuclear Non-Proliferation Treaty gives all signatories the right to develop a full-scale fuel cycle, and with it the technological infrastructure for making bombs. President Bush has begun to address this, by proposing that the spread of reprocessing technology be curtailed, with waste management and nuclear fuel supplied at reasonable cost — although to be acceptable, such a scheme should be run by an international entity such as the International Energy Agency or the International Atomic Energy Agency.

Properly managed, nuclear power can meet growing energy demand safely, cleanly and indefinitely.

Gerald E. Marsh*, George S. Stanford†

*5433 East View Park, Chicago, Illinois 60615, USA

†4700 Highland Avenue, Downers Grove, Illinois 60515, USA

BOOKS & ARTS

The world at your fingertips

The rise of the Internet search engine Google as guardian, gate-keeper and guide to a wealth of information.

The Search: How Google and its Rivals Rewrote the Rules of Business and Transformed Our Culture

by John Battelle

Portfolio/Nicholas Brealey: 2005. 320 pp.
\$25.95/£16.99

The Google Story: Inside the Hottest Business, Media and Technology Success of Our Time

by David Vise & Mark Malseed

Delacorte/Macmillan: 2005. 336 pp.
\$26/£14.99

Jon Kleinberg

Thirteen years ago, the release of a set of basic network protocols and browsing software ignited a combustible mixture of bandwidth, computing power and communication media. It was the beginning of the World Wide Web's dizzying rise from a small research project to a technological, economic and social force of global proportions.

Even in the early stages of the web's development, the notion of searching was seen as crucial, given the vast quantities of information available. But the true power of searching as an interface became apparent only more gradually, through the emergence of highly sophisticated tools such as Google, Yahoo and others. Searching is now our primary point of contact with the virtual world: we search to find out about people we are to meet, places we will visit, products we're about to buy; we search as a precursor to communicating and transacting business; and we search because it's the easiest way to get to almost anywhere we want in cyberspace.

The Search by John Battelle and *The Google Story* by David Vise and Mark Malseed are two recent books that develop this topic for a general audience. They adopt rather different approaches — Vise and Malseed focus primarily on Google, whereas Battelle ponders the idea of searching more generally — but each provides powerful illustrations of the extent to which searching has slipped free of its technological moorings and become a force in both the information economy and in everyday life.

Battelle's book is appealingly organized around a story and a big idea. The story is the history of web searching, with Google as its hero and dominant character; the big idea is what Battelle calls the "database of intentions", the archived queries and actions of users at



Sign of the times: Google co-founders Larry Page (left) and Sergey Brin have an icon on their hands.

sites such as Google, Yahoo, AOL, Amazon and eBay. The database of intentions is not simply a thought experiment designed for the purposes of the book; such archives exist among the massive data sets maintained by these companies. And they form a cultural record, capturing our era's collective interests and desires at a scale and resolution unprecedented in human history.

The story and the big idea are inextricably linked, of course. To begin with, the search industry became profitable, and hugely so, thanks to advertising targeted directly at the intentions of its users. For example, Google's AdWords service, based on a methodology introduced by its competitor Overture, displays ads that are targeted to the current query. This monetization of intent is the basis of the search economy — advertisers can pitch their messages to a user at precisely the moment the user has expressed a concretely formulated interest.

Battelle follows the many strands of this notion, including the tension between paid listings and the 'pure' algorithmic search results that are the staple of Google and other search engines; the businesses that live or die

based on traffic from search engines; and the new legal and economic terrain into which these developments have led us. Battelle also asks us to consider the dangers inherent in consolidating so much sensitive, highly personal information in the hands of any company — even one that, like Google, promises in its public statements not to be "evil". He offers illuminating discussions and hypothetical scenarios covering the many ways such data could potentially be used and misused, and this leads to an inescapable conclusion: if you aren't at least a little bit afraid of this future, you probably haven't thought about it hard enough.

Vise and Malseed address several of these economic, legal and social issues as well, but with more of an emphasis on Google specifically. This approach limits their consideration of certain broader directions, but it leads them more deeply than Battelle into other topics. Of particular interest is their emphasis on Google's stunning computational resources — a customized system built on more than 100,000 commodity-level personal computers — and on some of the future projects, including potential forays into bioinformatics and

B. MARGOT/AP PHOTO

genomics, that this technology makes possible.

Each book contains, as a core element, the story of Google itself, from the initial meeting of the founders, through its intellectual and technological evolution as a university research project, and on to its emergence as a public company, a verb and a potent icon of pop culture. It is a fascinating and compelling story, even for those who know its broad outlines.

And given the company's origins, it can be read as a parable on the value of fundamental research — on the way the pursuit of long-range, scientifically challenging goals can have pay-offs that extend to the public at large and can ultimately change the world. ■

Jon Kleinberg is in the Department of Computer Science, Cornell University, Ithaca, New York 14853, USA.

The bright side of life

Aglow in the Dark: The Revolutionary Science of Bioluminescence

by Vincent Pieribone & David F. Gruber

Belknap Press: 2006. 288 pp

\$24.95, £15.95, €23.10

Thomas G. Oertner

Creatures that glow in the dark have evolved independently many times in various branches of the phylogenetic tree, and it is safe to assume that it has always been for communication of one kind or another. Whether you are trying to lure prey at the bottom of the sea, startle or confuse attackers, keep your swarm together, or signal your availability to potential mates, being able to send out light in a controlled fashion will give you an edge, especially in deep-sea environments. But although the evolutionary value of bioluminescence seems obvious, the question of how it works has proved difficult to answer.

The first clue was found 120 years ago by the French physiologist Raphaël Dubois. From his experiments with the light organ of the beetle *Pyrophorus*, he concluded that it must contain a two-component system: a heat-sensitive catalyst that he termed luciferase, and a fuel component, luciferin. Purifying the components turned out to be quite difficult, however, and it was only 50 years ago that a young Japanese scientist, Osamu Shimomura, finally succeeded in producing crystals of pure luciferin.

A few years later, Shimomura managed to reveal the light-production machinery of the jellyfish *Aequorea*, in which luminescence is tightly controlled by the concentration of intracellular calcium. He also noted that the light emitted by aequorin molecules in a test tube is bright blue, whereas the jellyfish glows with a greenish hue. The green fluorescent protein (GFP) responsible for this energy conversion received little attention at the time but revolutionized biology 30 years later. Today, customized fluorescent proteins are used as reporters of gene activation and cell identity, to visualize the subcellular localization of tagged proteins, and to monitor cellular activity by tapping in to various second-messenger systems. The most spectacular emerging application is probably the visualization of nerve-cell activity in the brains of living animals.

Given its multitude of uses, I was surprised

to learn that the publication announcing the complete sequence of GFP (D. C. Prasher *et al.* *Gene* **111**, 229–233; 1992) was greeted by just two requests for the complementary DNA clone. The second clone, however, ended up in the hands of a man who not only immediately realized its scientific (and economic) potential, but also had the creativity and skills to transform it into the powerful tool it is today. Even if you can already guess his name (it was Roger Tsien), you will probably still enjoy *Aglow in the Dark*, Vincent Pieribone and David Gruber's well-narrated and beautifully illustrated book. It combines character studies of the people involved with a thoroughly researched story of the unlikely events that led to the main discoveries. The authors interviewed many of the key players, including the Russian scientists who were first to discover a red fluorescent protein. The book's journalistic style gives it a more 'real' feel than its forerunner *Glowing Genes* by Marc Zimmer (Prometheus, 2005), which relies mostly on secondary information.

Interspersed with the main narrative are chapters that explain basic concepts, including the nature of light and the genetic code. Although these excursions might help to reel in readers from other disciplines, in some cases the authors go off on a tangent to cover

subjects with no obvious relevance to the central theme, such as Schrödinger's cat. The authors' lengthy explanation about the brevity of laser pulses used for two-photon microscopy seems designed to 'wow' an unsophisticated audience with far-out facts. Not only do they get the calculation wrong, but why should we care about it anyway?

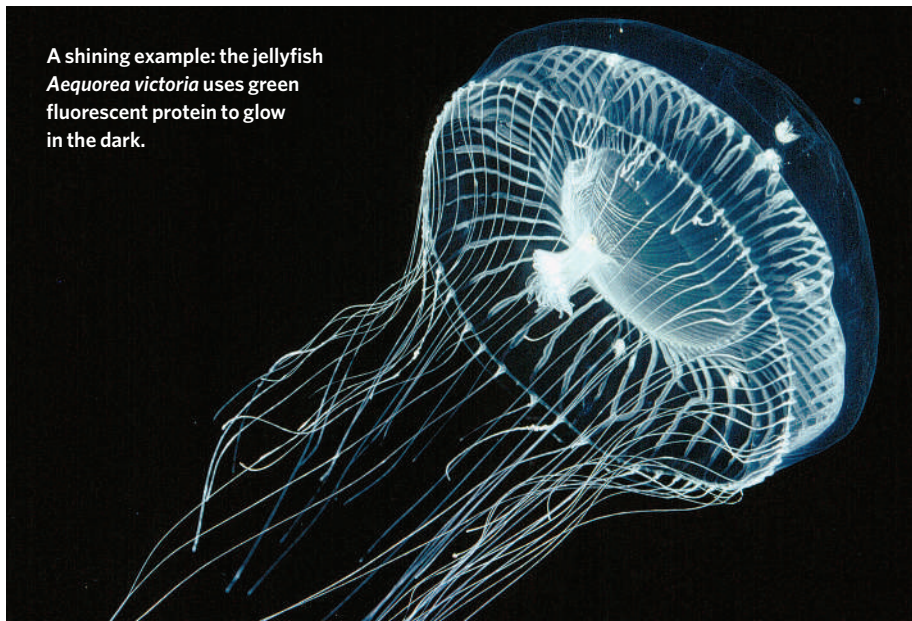
These quibbles aside, the main narrative is riveting, and the authors capture the sometimes curious way that science progresses through an alternation of chance discoveries and systematic, goal-directed experiments. Students wondering whether they are cut out to become scientists ought to be encouraged by the diverse cast of characters involved in solving the mystery of bioluminescence.

Obscure marine invertebrates have been a significant resource for discoveries of fundamental importance, from Alan Hodgkin and Andrew Huxley's first recording of action potentials in the squid to the highly specific neurotoxins that Baldomero Olivera found in Pacific cone snails. In this sense, the story of fluorescent proteins is another timely reminder of the value of biodiversity. The discovery in 2003 of light-activated ion channels from the photosensor of the green alga *Chlamydomonas* makes one wonder what else is out there waiting to be found and put to use. It also suggests that the optical brain-machine interface that the authors discuss in the final 'sci-fi' chapter of the book could soon become a reality, using light not only as a readout of activity, but also for the precise stimulation of individual nerve cells.

These are exciting times for biology, and this accessible and lively introduction conveys the sheer pleasure of discovery, as well as the enormous technological potential of fluorescent proteins. ■

Thomas G. Oertner is at the Friedrich Miescher Institute, Maulbeerstrasse 66, 4058 Basel, Switzerland.

A shining example: the jellyfish *Aequorea victoria* uses green fluorescent protein to glow in the dark.



C. MILLS



S. KLIPPER

The Ross ice shelf in Antarctica covers an area of half-a-million square kilometres.

Journey to the end of the Earth

Terra Antarctica: Looking Into the Emptiest Continent

by William L. Fox

Trinity University Press: 2005. 304 pp. \$35

Lloyd Peck

Given the inaccessibility and mystery of Antarctica, it is not difficult to see why the continent inspired writers such as Edgar Allan Poe and Jules Verne to set novels in a fictitious ice-bound land with hot plateaux where dinosaurs persist, and deep craters that provide access to the centre of the Earth. Antarctica certainly poses a challenge for the imagination of anyone who contemplates its vast beauty. In *Terra Antarctica*, William Fox aims to show how we use cultural means to augment our neurobiology in overcoming the perceptual difficulties we experience when exploring this huge, largely deserted continent.

The book itself is an eclectic journey. It melds a series of deeply intense personal experiences in Antarctica with well-researched vignettes on the history or development of artistic or cartographic representation. This is then put into the context of our current view of Antarctica and of the science done there. This can, at times, seem like a bizarre combination. Juxtaposed with awe-inspiring descriptions of the landscape are, for example, sections on the degree angular scale interferometer, which measures microwave background radiation and provides insight into the

shape of the Universe and what it was like in its early stages.

Polynesian history is also included, as the Polynesians are thought to have been the first to discover the Antarctic. According to their oral history, Ui-te-Rangiora sailed south to a place where the sea was white and dangerous, a voyage that probably took place around AD 650, more than 1,000 years before Western civilization found the continent. This then contrasts with a detailed description of the influence of the painter J. M. W. Turner on the representation of landscape in art and the development of photographic methods to document landscape. This style of presentation may not be to everyone's taste.

Clearly, drawing together such a great diversity of subjects is more than a challenge. Fox does it well, making good use of his Antarctic experiences, and there is much genuine awe in his responses during his visits to research sites at McMurdo and Pole stations, and also to the driest deserts on Earth and the most southerly active volcano, Mount Erebus.

The historic elements, the descriptions of how the methodology developed, and the detailed science concepts are written in a clear and accessible manner that will bring the subjects into deep relief for many. The direction of the plot is more erratic than in books on a set science or arts subject, however. There are also some enthralling impressions of a landscape that is clearly a challenge to encompass, even

for someone with Fox's powerful ability to paint verbal pictures. I liked his description of the seemingly endless variety of ice in Lake Vanda that Fox claims left him dumbstruck (a common response to Antarctica's ability to go beyond the human imagination) as he stood surrounded by acres of glass-clear ice fractured into amazingly intricate geometries. This section is written almost as a religious experience in which, standing on the ice, the feeling is of flying in a dream. Descriptions of glass-corroding acid clouds on Mount Erebus, and tent-toppling storms with 85-mile-an-hour winds that hurl golf-ball-sized pebbles at you, are more the stuff of Antarctic lore, but they are still gripping and fantastical to those who have not seen them.

Readers interested in how humans have struggled with nature to develop techniques to master the inaccessibility of the world around them will want this book on their shelves. Those whose feet are more firmly entrenched in the real world, and who eschew a metaphysical approach to understanding the world, may be frustrated by the eclectic nature of the content. Those in between will find some things they like or cherish, and some that pass them by.

In the end, Antarctica is too much for even the greatest wordsmiths' use of language, the greatest painters' palettes, and the greatest photographers' use of light and composition. Fox gives us an enthralling guided tour of the human mind's attempt to make space into place, and land into landscape. But for me, at least, Antarctica remains beyond the reach of human arts.

Lloyd Peck is at the British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, UK.

NEWS & VIEWS

NANOSTRUCTURES

The manifold faces of DNA

Lloyd M. Smith

When it comes to making shapes out of DNA, the material is there, and its properties are understood. What was missing was a convincing, universal design scheme to allow our capabilities to unfold to the full.

As civilization has developed over the past 10,000 years, humankind has learned how to build larger and larger structures; over the past two decades, we have begun to learn how to build smaller and smaller structures. On page 297 of this issue¹, Paul Rothemund presents a material step forward in this second arena: he describes a stunningly simple and versatile approach to the fabrication, by self-assembly, of two-dimensional DNA nanostructures of arbitrary shape.

DNA has emerged over recent years as the molecule of choice for nanodesigners. There are two reasons for this. First, in the 50 years since the discovery of the DNA double helix², a detailed understanding of the energetics of its formation has developed³. This allows one to predict, with reasonable success, the shapes into which a DNA molecule of a given sequence will fold in solution⁴. Second, the advent of automated chemistry for the rapid synthesis of DNA molecules has made it possible to easily obtain DNA molecules of any desired sequence, of lengths up to 100 nucleotides or so. (A nucleotide is the monomer unit of DNA.)

Longer molecules may also be obtained using biological methods such as the polymerase chain reaction or now-standard molecular cloning approaches. The molecular engineer is thus armed with two of the basic elements needed to build structures of interest: the materials for building and an understanding of their properties. The missing ingredient has been a versatile design strategy.

The best-developed model for DNA design is the 'tile model' developed in Ned Seeman's laboratory⁵. This uses as its basic building-block DNA in two-dimensional, rectangular shapes. These tiles are designed to include crossover points between DNA strands, so imparting stiffness to the structure. Free single-stranded regions ('sticky ends' in the parlance of the molecular biologist) extrude from each corner of the tile and permit tiles to self-assemble into larger, two-dimensional sheets.

This basic concept has been adapted and extended many times, enabling demonstrations of, for instance, the templated self-assembly of protein arrays⁶, and the

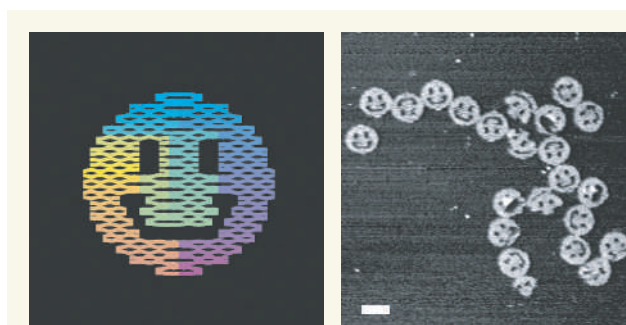


Figure 1 | Give us a smile.
a, A folding path to make a DNA 'smiley'. An even number of double helices is filled into the desired shape in the order of colours in the spectrum from red (start) to violet (end).
b, An atomic force microscopy image of the finished product. (From Fig. 2 on page 299; scale bar, 100 nm.)

fabrication of the fractal patterns known as Sierpinski triangles⁷.

A second design theme was introduced by William Shih and colleagues⁸ in 2004. They showed that a single strand of DNA, 1,669 nucleotides long, could be driven to self-assemble into a nanoscale octahedron by the addition of five short DNA strands complementary to selected regions of the original strand. Again, structural rigidity was obtained by means of crossover points. This work was notable in two respects. First, it permitted assembly of a three-dimensional structure, rather than the two-dimensional sheets that had been the primary focus of most previous work. Second, it introduced the concept of using short DNA strands to direct the folding of a longer DNA strand.

Rothemund¹ builds upon both these models, and expands their generality to permit the fabrication in DNA of any two-dimensional shape. Just as in the approach described by Shih *et al.*⁸, he directs the folding of a long single-stranded DNA molecule, in this case the genome of the widely used cloning vector M13 mp18. M13 is a bacteria-destroying virus with a single-stranded DNA genome about 7,000 nucleotides in length; its known sequence and ready availability make it convenient for this application.

The design process¹ has five steps (see Fig. 1 on page 298). First, the desired shape is chosen and is filled from top to bottom by an even number of parallel double helices. Second, a single, long scaffold strand is folded back and forth along the double helices, so introducing periodic crossovers (again for rigidity)

between parallel helices. Third, a computer program generates the sequences of many short 'staple strands'. These bind to the DNA scaffold strand — making it double-stranded — and create crossovers between strands. In the two final steps, the design is examined and refined by computer to relieve strain and to strengthen the structure at the nicks and seams produced in the initial design process.

The results that emerge are stunning. Rothemund shows the generality of the approach with six different structures (Fig. 2 on page 299), notably a five-pointed star and a smiley face — myriads of which are a disconcerting sight in an atomic force microscopy image (Fig. 1, above). He further demonstrates the assembly of the individual structures into rather beautiful, higher-order patterns, reminiscent of the designs found in Persian carpets, and shows the absence of any symmetry requirement in the designs by fabricating a map of the world (Fig. 3g on page 300).

Rothemund's basic method is fairly straightforward, and ample experimental and design details are provided in the many pages of supplementary material. He notes that there is a plethora of widely available chemical modifications to DNA strands; this should make it possible to incorporate, for example, dyes or binding elements into these structures at any desired position. The extension of this model from two to three dimensions should not prove too difficult either, given the three-dimensional precedents⁸, opening up further possibilities in the design and construction of functional materials at the nanoscale.

Thus equipped not only with DNA building

materials and an understanding of their structural and chemical properties, but also with a versatile general approach to weaving them together¹, we are arriving at a new frontier in our pursuit of ever-smaller structures. The barrier we have to surmount next is to deploy our knowledge to develop structures and devices that are really useful. Happily, in that endeavour we are now perhaps limited more by our imagination than by our ability. ■

Lloyd M. Smith is in the Department of Chemistry, University of Wisconsin, 1101 University Avenue,

Madison, Wisconsin 53706, USA.

e-mail: smith@chem.wisc.edu

1. Rothmund, P. W. K. *Nature* **440**, 297–302 (2006).
2. Watson, J. D. & Crick, F. H. *Nature* **171**, 737–738 (1953).
3. Breslauer, K. J., Frank, R., Blocker, H. & Marky, L. A. *Proc. Natl Acad. Sci. USA* **83**, 3746–3750 (1986).
4. Zuker, M. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
5. Seeman, N. C. *Nature* **421**, 427–431 (2003).
6. Yan, H., Park, S. H., Finkelstein, G., Reif, J. H. & LaBean, T. H. *Science* **301**, 1882–1884 (2003).
7. Rothmund, P., Papadakis, N. & Winfree, E. *PLoS Biol.* **2**, e424 (2004).
8. Shih, W., Quispe, J. & Joyce, G. *Nature* **427**, 618–621 (2004).

ALZHEIMER'S DISEASE

A needle from the haystack

Richard Morris and Lennart Mucke

Abnormal protein clumps of many varieties build up in the brains of individuals with Alzheimer's disease. But which types actually cause memory deficits? The behaviour of model mice might help to find out.

A huge array of neurodegenerative diseases seems to be caused by abnormal clusters of certain proteins. Many of these disorders are on the rise and cannot be treated effectively, including the most frequent among them — Alzheimer's disease¹. To understand and treat these conditions better, much more needs to be learned about the structures and activities of the disease-causing protein assemblies. In this issue, Karen Ashe and her colleagues (Lesne *et al.*, page 352)² describe the isolation of a specific protein assembly from mouse brains that may help to explain some of the memory loss seen in Alzheimer's disease. Remarkably, they applied a behavioural screen to select the assemblies that were functionally most relevant.

The key component of Lesne and colleagues' protein assembly is the amyloid- β peptide ($A\beta$), which is widely believed to have a causal role in Alzheimer's disease. Compelling support for this $A\beta$ hypothesis comes from studies indicating that most, if not all, genetic causes and risk factors for Alzheimer's increase the cerebral accumulation of $A\beta$ (ref. 3), and from a variety of experimental models in which $A\beta$ impairs neural functions^{4,5}. However, it is still uncertain how $A\beta$ accumulation may lead to the relentless cognitive decline that characterizes the disease^{4,6}.

A major source of confusion has been the propensity of $A\beta$ to exist in a variety of complexes that seem to differ not only in their number of peptide building-blocks, but also in their overall conformation and biological activity (Fig. 1). In increasing order of

complexity, $A\beta$ can exist as monomers (a single peptide unit each), dimers (pairs), trimers (trios), oligomers (many units), tiny transient structures known as protofibrils, larger stable fibrils, and highly compacted admixtures of fibrils and smaller aggregates (amyloid plaques). To make matters worse, these different assemblies can be bound to a variety of other structures within the complex molecular and cellular environment of the

brain. It is therefore not surprising that it has taken a while to determine which assembly is most closely related to cognitive decline.

Building on their own studies and pioneering work of other groups^{4,6}, Lesne *et al.* went on a biochemical hunt for this $A\beta$ assembly in a mouse model for Alzheimer's disease. Although it is, of course, impossible to recapitulate all aspects of this complex human disease in a mouse, mice genetically engineered to express human amyloid precursor protein (APP) and the APP-derived $A\beta$ peptides in their neurons develop a highly informative array of Alzheimer-like alterations, including cerebral amyloid plaques, distortions of neuronal branches, impairments of synapses (the junctions between neurons) and deficits in learning and memory⁵.

Rather than work all the way back from end-state disease to altered physiology, the authors used the onset of early memory deficits in APP mice as an instructive guide in their detection of the $A\beta$ assemblies that are most likely to impair neuronal function. Spatial memory impairments in these animals occurred in discrete stages. APP mice remembered as well as normal control animals when they were young. But at 6 months of age, there was a partial decline in memory to a point that then remained stable for 7–8 months. This was followed by a further memory decline at around 15 months. Lesne and colleagues used this graded pattern of forgetfulness as their route to trapping the molecular culprit.

They combined their behavioural studies with a detailed biochemical analysis of the $A\beta$ assemblies found in the brains of the animals. Previous studies had already shown that plaque load and overall $A\beta$ content are poor predictors of cognitive failure^{4,5,7}. A series of tissue fractionation and protein purification steps allowed the researchers to zero in on a specific, functionally relevant $A\beta$ assembly, which they termed $A\beta^{*56}$. This name arose because the assembly reacted with anti- $A\beta$ antibodies, tandem mass spectrometry showed it contained $A\beta$ sequence, and it had an apparent relative molecular mass of 56,000 in gel electrophoresis. The authors provide evidence that $A\beta^{*56}$ forms outside cells and that it may be a cluster of 12 $A\beta$ peptides — although the exact origin and atomic structure of this assembly remain to be resolved, and it may be that it has additional constituents other than $A\beta$.

The early memory deficits in APP mice, starting at 6 months, were negatively correlated with the level of $A\beta^{*56}$ found in their brains. In more direct support of a causal relationship, an $A\beta^{*56}$ -containing isolate from APP mouse

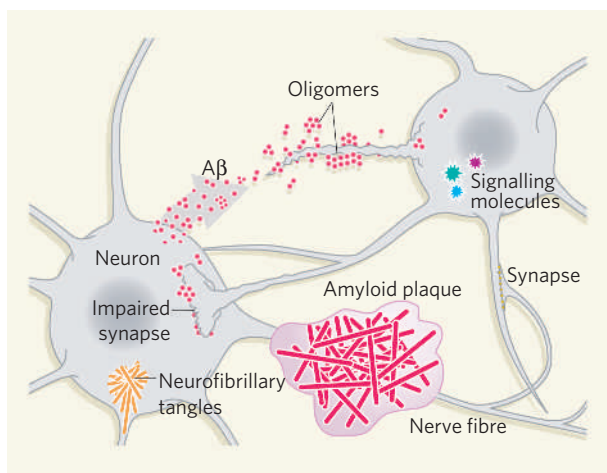


Figure 1 | The role of amyloid- β ($A\beta$) in Alzheimer's disease. $A\beta$ peptides self-aggregate and form increasingly larger assemblies that accumulate in the brain. The largest of these assemblies (amyloid plaques), a hallmark of Alzheimer's disease, displace and distort neuronal branches. Smaller $A\beta$ assemblies (oligomers) have been much harder to detect in brain tissues but may be even more detrimental than plaques. Lesne *et al.*² show that $A\beta^{*56}$ is among these stealthy moieties, and seems to have a key role in memory decline. Independent of neurofibrillary tangles — another pathological hallmark of the human condition — $A\beta$ oligomers may impair the junctions through which brain cells communicate (the synapses) or alter neurotransmission by other mechanisms. This would deplete signalling molecules that are dependent on synaptic activity and are required for memory and other brain functions.

CHEMICAL BIOLOGY

A pocketful of colour

One of the joys of summer evenings is watching fireflies light up the night sky. These twinkling bugs communicate with one another using flashes of light — a trick biologists have appropriated to convey readouts from their experiments. Various forms of the enzyme responsible for the fireflies' luminescence — luciferase — can produce slightly different colours, ranging from red to yellow to green. Elsewhere in this issue, Toru Nakatsu *et al.* (*Nature* **440**, 372–376; 2006) report several X-ray crystal structures of luciferase from the Japanese Genji-botaru firefly (*Luciola cruciata*, pictured) that explain how small changes in this protein can change the colour of the emitted light.

To create bioluminescence, magnesium, adenosine triphosphate and a small molecule called luciferin react with molecular oxygen. This reaction, which is catalysed by luciferase, generates an electronically excited oxyluciferin species. The yellow-green sparks we see in the night sky are emitted when oxyluciferin relaxes from its excited state to the ground state, losing energy in the form of light.

Single amino-acid changes in the active site of luciferase can alter the colour of the light emitted by the protein, but the chemical mechanism involved has been a mystery. Now Nakatsu *et al.* have obtained a series of 'snapshots' of the reaction — luciferase bound to the reactants (adenosine triphosphate and magnesium); an analogue of one of the reaction intermediates (known as DLSA); and the products (oxyluciferin and adenosine monophosphate).

The structure of luciferin bound to the reactants and the structure of the products are similar, both possessing an active site that is 'open' to the environment surrounding the protein. But when the protein is bound to DLSA, the active site closes: DLSA is tightly packed against several amino-acid side-chains, including isoleucine 288. Analysis of the three structures revealed that the side-chains of isoleucine 288 and serine 286 rotate when DLSA is present, closing the active site and forming a pocket.

The authors thought that the conformation of these amino acids might affect the colour of the light



S. KURIBAYASHI

discharged by luciferase. So they solved the X-ray crystal structure of DLSA bound to a mutant luciferase known to emit red light. Their theory was borne out, as the active site of the DLSA-bound mutant protein was open.

The authors propose that the colour of the emitted light depends on the formation of this compact microenvironment during the reaction. In the normal protein, the excited oxyluciferin is held tightly in this pocket, packed against the side-chain of isoleucine, and yellow-green light is emitted. However, if this rigid pocket does not form — for example,

in proteins that have mutations at isoleucine 288 or serine 286 — the excited oxyluciferin loses some energy, possibly because it can move around a little, and the protein emits red light, which is lower in energy.

In further work, Nakatsu *et al.* found that many of the other firefly luciferases that emit yellow-green light have either isoleucine or the slightly smaller amino acid valine at position 288. So it seems likely that the movement of the 288 amino acid is a common mechanism for controlling the colour of luciferase bioluminescence.

Joshua Finkelstein

brain elicited new memory deficits when injected into the brains of rats (rather than mice). The rats had learned one spatial task already (the water maze), but then had to learn another with A β *56 permeating their forebrain. Learning seemed to proceed rapidly but, when tested a day later, the A β *56-treated animals failed to show effective spatial memory. Animals injected with a control solution had no trouble remembering the second task. A β *56 did not, however, cause lasting cognitive problems, as both groups of rats were just as good as each other in learning a third task when trained 10 days later (there was no further injection of A β *56). These data imply that A β *56 causes a transient disruption of the physiological mechanisms responsible for memory, rather than permanent neuronal damage.

During the second decline in memory (beginning around 15 months), the A β *56 in the APP mouse brains did not rise. This might be viewed as inconsistent with a causal role of this A β assembly in memory decline. However, the dissociation could be explained by compensatory mechanisms that were able to counteract the pathogenic effects of A β *56 for a while but then failed as a consequence

of old age. Alternatively, different stages of A β -induced neurological disease might involve distinct pathogenic mechanisms.

As with many discoveries, the findings by Lesne *et al.* raise questions. Can A β *56 be detected in other APP models with memory deficits? What is the relation between A β *56 and the A β oligomers detected in the cerebrospinal fluid of Alzheimer's patients⁸? Are these A β assemblies a reliable biomarker for early diagnosis, monitoring disease progression and evaluating new treatments? Might these A β assemblies themselves constitute targets for more effective therapeutic interventions^{4,6}? And last, but not least, how do they actually derange neuronal function?

In regards to this last question, it is noteworthy that abnormalities in signalling through neuronal glutamate receptors have been identified as a likely mechanism of A β -induced neuronal and behavioural deficits^{7,9}. This is an intriguing possibility because hippocampal glutamate receptors have a key role in spatial learning and memory¹⁰. It will be interesting to determine whether A β *56 has an effect on these receptors or related intracellular signalling pathways. A β *56 is certainly a star suspect, but there

is more detective work to be done before it can be convicted as the real culprit. ■

Richard Morris is in the Laboratory for Cognitive Neuroscience, College of Medicine and Veterinary Medicine, University of Edinburgh, 1 George Square, Edinburgh EH8 9JZ, UK.

e-mail: r.g.morris@ed.ac.uk

Lennart Mucke is at the Gladstone Institute of Neurological Disease and the University of California, San Francisco, 1650 Owens Street, San Francisco, California 94158, USA.

e-mail: lmucke@gladstone.ucsf.edu

1. Hebert, L., Scherr, P., Bienias, J., Bennett, D. & Evans, D. *Arch. Neurol.* **60**, 1119–1122 (2003).
2. Lesne, S. *et al.* *Nature* **440**, 352–357 (2006).
3. Tanzi, R. & Bertram, L. *Cell* **120**, 545–555 (2005).
4. Walsh, D. M. & Selkoe, D. J. *Neuron* **44**, 181–193 (2004).
5. Kobayashi, D. T. & Chen, K. S. *Genes Brain Behav.* **4**, 173–196 (2005).
6. Klein, W. L., Krafft, G. A. & Finch, C. E. *Trends Neurosci.* **24**, 219–224 (2001).
7. Palop, J. J. *et al.* *J. Neurosci.* **25**, 9686–9703 (2005).
8. Georganopoulou, D. G. *et al.* *Proc. Natl Acad. Sci. USA* **102**, 2273–2276 (2005).
9. Snyder, E. M. *et al.* *Nature Neurosci.* **8**, 1051–1058 (2005).
10. Morris, R. G. M., Anderson, E., Lynch, G. S. & Baudry, M. S. *Nature* **319**, 774–776 (1986).



50 YEARS AGO

"The Arrow of Time"

— It is widely believed that all irreversible mechanical processes involve an increase of entropy, and that 'classical' (that is, non-statistical) mechanics, of continuous media as well as of particles, can describe physical processes only in so far as they are reversible in time. This means that a film taken of a classical process should be reversible, in the sense that, if put into a projector with the last picture first, it should again yield a possible classical process. This is a myth, however, as a trivial counter-example will show. Suppose a film is taken of a large surface of water initially at rest into which a stone is dropped. The reversed film will show contracting circular waves of increasing amplitude. Moreover, immediately behind the highest wave crest, a circular region of undisturbed water will close in towards the centre. This cannot be regarded as a possible classical process. (It would demand a vast number of distant coherent generators of waves the co-ordination of which, to be explicable, would have to be shown, in the film, as originating from one centre. This, however, raises precisely the same difficulty again, if we try to reverse the amended film.)

K. R. Popper

From *Nature* 17 March 1956.

100 YEARS AGO

Journals dealing with the chemical aspects of physiological and pathological research have long been current in Germany; but up to the present time English-speaking workers have had to rely on periodicals dealing with all branches of physiology and pathology for the publication of their results... the need has for long been felt of a special journal, and we have to chronicle the advent of one — the *Bio-chemical Journal*... In America also a similar want has been met by the issue of the first numbers of what is there called the *Journal of Biological Chemistry*.

From *Nature* 15 March 1906.

PHOTOCHEMISTRY

Lighting up nanomachines

Euan R. Kay and David A. Leigh

A cleverly engineered molecule uses light to generate a charge-separated state and so cause one of its components to move. It's the latest study of a molecular machine that exploits nature's most plentiful energy source.

Nature runs the nanomachinery that makes life possible using the last word in clean, free and readily available power sources — sunlight. In photosynthetic bacteria and green plants, photon absorption by chlorophyll generates a charge-separated state, from which the electron is quickly passed down a cascade of electron carriers, ultimately generating energy in a convenient chemical form. Can similar capabilities be engineered? An exemplary effort to do just this is given by Balzani *et al.* who, writing in *Proceedings of the National Academy of Sciences*¹, describe photochemical experiments on an artificial machine that uses light to displace a fragment of its unimolecular structure.

Those who seek to harness the Sun's energy for synthetic molecular machines find that chemistry is always throwing up obstacles. In particular, charge recombination typically occurs thousands or millions of times faster than the nuclear movements on which such machines rely, making charge-separated states difficult to exploit. This problem can be overcome using bimolecular systems: here, the charged partners quickly diffuse apart so their energy can be used, for example, to achieve

switching in a rotaxane². This class of molecule, consisting of a ring that shuttles randomly and incessantly along a string, stopped only by bulky groups at the string's termini, is also that used by Balzani and colleagues¹.

Their rotaxane³ (Fig. 1) incorporates two structurally different bipyridinium sites — 'stations' 1 and 2 — that slow the shuttling ring's motion through strong short-range electrostatic interactions. The ring thus divides its time between station 1, station 2 and the rest of the string in the ratio of around 95:5:1. At room temperature, the ring shuttles between the stations tens of thousands of times per second, but the net flux is zero. So no work can be done, or useful task performed, by the shuttling action (the 'principle of detailed balance'⁴).

One of the bulky end-groups of the rotaxane's string is a ruthenium trisbipyridine complex. This can absorb a photon of visible light and so form a reactive, excited state that donates an electron to the more easily reduced of the two bipyridinium sites — station 1, the ring's preferred binding site. One would normally expect the resulting charge imbalance to be corrected by back-transfer of an electron on

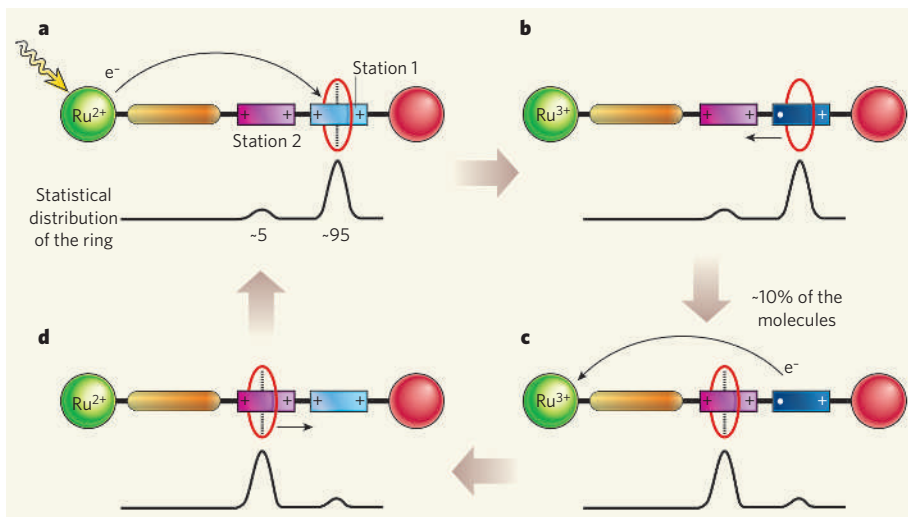


Figure 1 | Light-driven molecular shuttle. Balzani and colleagues' rotaxane^{1,3} consists of a molecular ring free to move along a molecular string. **a**, At equilibrium in the ground state, the ring spends most of the time over station 1, as a result of attractive, non-covalent interactions. But irradiation of the ruthenium complex (green) at one end of the string generates a highly reducing excited state, resulting in electron transfer to station 1, and the weakening of this station's electrostatic interactions with the ring. **b**, Normally, charge recombination is fast in comparison with nuclear motions, but here a delay allows approximately 10% of the molecules to undergo significant brownian motion, shifting the distribution of these rings to favour station 2. **c**, When charge recombination eventually does take place, the higher binding affinity of station 1 is restored, and **d**, the system relaxes to restore the original statistical distribution of rings.

the sub-microsecond timescale, even given the considerable distance between the ruthenium complex and the stations (which also slows down the rate of the forward electron-transfer reaction). In fact, the charge-separated state has an average lifetime of around 10 microseconds.

Reduced bipyridinium is a very much poorer binding site for the ring, and the charge-separated state survives long enough for about 10% of the rings to undergo significant brownian motion. Detailed balance is broken and a net flux of rings occurs as they shift their allegiance to the unreduced station 2. After 10 microseconds, however, as back electron transfer finally takes place, station 1 regains its stickiness. A net flux of rings occurs back from station 2, the system returns to its original equilibrium, and a machine cycle has taken place. The process can be repeated for at least a thousand cycles.

This is a fascinating system, working (as do most other light-driven shuttles^{2,5–8}) without the consumption of chemical fuels or the formation of waste products. It can be properly called a machine because component displacements occur in response to an external stimulus. But is it really best described as a 'nanomotor'? Chemists are still pondering the most useful way to understand the behaviour of the contemporary, early generations of synthetic molecular machines, but physics and biology offer many useful phenomenon-based ideas in this regard. Brownian motors, for example, use mechanisms⁹ that harness random molecular-level motion like that of the ring motion in rotaxanes.

Such brownian ratchet mechanisms, which are believed to account for the behaviour of some motor proteins¹⁰, all require detailed balance to be broken to allow a net, directed flux of particles. Crucially, however, the 'ratchet' part of the mechanism ensures that the resulting change in particle distribution is not undone when the motor is reset. This allows the machine to be able to pump the particle distribution further and further away from equilibrium (as with the enzymes that synthesize the currency of intracellular energy transfer, ATP) or move itself progressively down a track (as with the cell's internal pack-horse, kinesin).

This feature is missing from Balzani and colleagues' rotaxane¹, and other simple molecular shuttles^{2,5–8}: the work done in breaking detailed balance is undone by the reset step. For such a system to be understood to be a motor by a statistical physicist or biologist, therefore, the rings would have to be diverted along a different track during the reset phase (making a rotary motor), or remain where they are while the machine is reset (making a linear motor or pump). The former is the basis of several wholly synthetic molecular motors^{11–15}; the latter has thus far been achieved only with artificial structures made from DNA¹⁶.

Balzani and colleagues' shuttle operates

through a fully autonomous photochemical cycle; but can these molecules repetitively do work as long as sunlight is available? The authors did not use sunlight in the experiments they report, but instead treated the rotaxanes with a single 10-nanosecond laser pulse at a visible-light wavelength. If continuously irradiated with sunlight, the distribution of the rings between the two stations would reach a steady state (the exact distribution depending on the intensity of the light) within a few milliseconds. To generate net flows of rings between the stations after that, it appears one would have to switch the light source rapidly on and off.

This is in contrast to the performance of another family of machine molecules that have components that rotate directionally, rather than shuttle linearly^{12,17}. When these rotary molecules reach a steady state under continuous irradiation, as with the rotaxanes the bulk distribution of the machine components no longer changes. But even at the steady state, at a sufficiently high temperature, net fluxes of the rotor components still occur through different pathways between four different rotary isomers that are present. Under constant irradiation, the molecules thus operate continuously as directionally rotating motors¹⁷.

Synthetic molecular motors and machines are very much in their infancy, and chemists are still learning the most basic rules for their design and operation. It is a field that can usefully draw on input from physicists, biologists, engineers and materials scientists. For a deeper understanding of molecular machine systems to evolve, therefore, it would be highly beneficial if the terminology used to describe

them were to become consistent across all the contributing disciplines. Balzani and colleagues' latest photochemical experiments¹ represent a fascinating advance in our understanding of how a charge-separated state can be used to bring about a nuclear displacement in a unimolecular machine. It will doubtless prove an important stage on the route towards autonomous artificial nanomotors powered by sunlight.

Euan R. Kay and David A. Leigh are in the School of Chemistry, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JJ, UK.
e-mail: david.leigh@ed.ac.uk

1. Balzani, V. *et al. Proc. Natl Acad. Sci. USA* **103**, 1178–1183 (2006).
2. Brouwer, A. M. *et al. Science* **291**, 2124–2128 (2001).
3. Ashton, P. R. *et al. Chem. Eur. J.* **6**, 3558–3574 (2000).
4. Onsager, L. *Phys. Rev.* **37**, 405–426 (1931).
5. Murakami, H., Kawabuchi, A., Kotoo, K., Kunitake, M. & Nakashima, N. *J. Am. Chem. Soc.* **119**, 7605–7606 (1997).
6. Wurpel, G. W. H., Brouwer, A. M., van Stokkum, I. H. M., Farran, A. & Leigh, D. A. *J. Am. Chem. Soc.* **123**, 11327–11328 (2001).
7. Altieri, A. *et al. Angew. Chem. Int. Edn* **42**, 2296–2300 (2003).
8. Wang, Q.-C., Qu, D.-H., Ren, J., Chen, K. & Tian, H. *Angew. Chem. Int. Edn* **43**, 2661–2665 (2004).
9. Reimann, P. *Phys. Rep.* **361**, 57–265 (2002).
10. Astumian, R. D. *Science* **276**, 917–922 (1997).
11. Kelly, T. R., De Silva, H. & Silva, R. A. *Nature* **401**, 150–152 (1999).
12. Koumura, N., Zijlstra, R. W. J., van Delden, R. A., Harada, N. & Feringa, B. L. *Nature* **401**, 152–155 (1999).
13. Leigh, D. A., Wong, J. K. Y., Dehez, F. & Zerbetto, F. *Nature* **424**, 174–179 (2003).
14. Hernández, J. V., Kay, E. R. & Leigh, D. A. *Science* **306**, 1532–1537 (2004).
15. Fletcher, S. P., Dumur, F., Pollard, M. M. & Feringa, B. L. *Science* **310**, 80–82 (2005).
16. Kelly, T. R. *Angew. Chem. Int. Edn* **44**, 4124–4127 (2005).
17. Feringa, B. L., Koumura, N., van Delden, R. A. & ter Wiel, M. K. J. *Appl. Phys. A* **75**, 301–308 (2002).

PALAEONTOLOGY

Scales, feathers and dinosaurs

Xing Xu

A fossil dinosaur that 'nests' with feathered relations in the dinosaur phylogenetic tree did not, it seems, have feathers. The discovery will encourage a re-evaluation of feather evolution.

Only birds have feathers — or so we thought until the discovery of fossils of feathered dinosaurs in China and elsewhere¹. Since then, palaeontologists and biologists have together been painting a simple picture of feather evolution based on evidence from fossils and from developmental biology. A new example of a predatory dinosaur, described by Göhlich and Chiappe on page 329 of this issue², makes that picture a little bit more complicated.

The context is provided by Figure 1 (overleaf), which shows the generally understood relationships among the main groups of dinosaur (Aves — the birds — is the only one of these groups to have extant members).

Feathers are thought to be characteristic of the Coelurosauria, a group of theropod dinosaurs that includes Aves and several other groups. Because feathers are unique and therefore innovative in developmental and evolutionary terms³, we can infer their presence in extinct forms using a method called phylogenetic bracketing. Based on the presence of feathers in some extinct coelurosaurs and all living birds, this approach suggests that all coelurosaurs, with the possible exception of gigantic species such as *Tyrannosaurus rex*, are feathered⁴.

The small coelurosaur described by Göhlich and Chiappe now enters the picture. This new

dinosaur, a carnivore named *Juravenator*, is represented by a beautifully preserved skeleton collected from Late Jurassic deposits from Solnhofen, Germany — the source of all specimens of the earliest-known bird *Archaeopteryx*. At 151 million to 152 million years old, the deposits in which *Juravenator* was found are only 2 million to 3 million years older than those containing *Archaeopteryx*.

Given the worldwide rarity of complete specimens of small theropods from the Jurassic, the exceptionally well-preserved skeleton of *Juravenator* is in itself a notable find. Most significantly, however, the specimen preserves scaled skin around the tail and hindlimbs. This is a big surprise considering that *Juravenator* is a small coelurosaur and a member of a small group called the Compsognathidae — which includes the first known feathered dinosaur, *Sinosauropteryx*, the tail and hindlimbs of which were feathered⁵.

The evolution of biological structures must be studied within an evolutionary framework. In the case of feathers, a robust theropod phylogeny is the basis for reconstructing the sequence in which feathers evolved. The distribution of various feather morphologies on the currently accepted phylogeny^{6–9} suggests that simple, filamentous feathers first evolved no later than the earliest stage of coelurosaurian evolution. More complex feathers with a thick central shaft and rigid symmetrical vanes on either side appeared early in the evolution of the coelurosaurian group Maniraptora; and feathers with aerodynamic features, such as a curved shaft and asymmetrical vanes, appeared within the maniraptors but before the origin of birds⁴. This inferred sequence of events is supported independently by developmental data^{3,10}. Göhlich and Chiappe² place *Juravenator* within the Compsognathidae, a group that is 'basal' in the coelurosaurian tree (Fig. 1). So *Juravenator* should bear filamentous feathers. But it seems to be a scaled animal, at least on the tail and hind legs.

Why, then, does a member of a feathered dinosaur family bear scales? The authors' answer² is straightforward: feather evolution, they say, is more complex than we thought. This could well be so. The early evolution of some major structures — such as the incorporation of jaw suspensory bones into the middle ear of mammals — is often flexible and experimental. It would not be surprising if feathers were lost and scaly skin re-evolved in some basal coelurosaurian species, or if feathers evolved several times independently early in

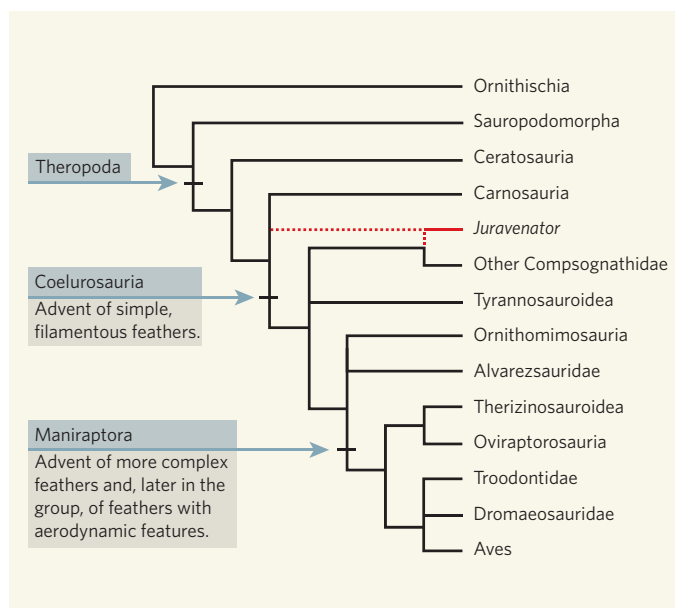


Figure 1 | Dinosaurs and feather evolution. This evolutionary tree of dinosaur groups is compiled from several phylogenetic analyses^{6–9}, and shows the possible occurrence of the main events in feather evolution. The tree shows alternative positions for *Juravenator*, the new species described by Göhlich and Chiappe², among the theropods. A compsognathid affinity for *Juravenator* suggests that feathers evolved independently or were lost in some species early in coelurosaurian evolution, or that some coelurosaurs had limited feathery covering. If *Juravenator* turns out to be more 'basal' in the tree than the known feathered dinosaurs, it would suggest that feather evolution started during the early history of the Coelurosauria.

coelurosaurian evolution. Or perhaps, suggest Göhlich and Chiappe, the absence of feathers on body parts that are feathered in other coelurosaurs is due to variability in the extent of a feathery covering among coelurosaurs. Although modern birds are extensively feathered, this may not have been true of their remote extinct relatives.

Feather preservation is extremely rare and biased. For example, contour feathers — those feathers forming the general covering of a bird — are not seen in some otherwise beautifully preserved specimens of *Archaeopteryx*, but are evident in other examples¹¹. Even after the discoveries of so many fossils of feathered dinosaurs, we still have only a patchy picture of feather distribution in basal coelurosaurs. And many details are missing — for example, we don't know whether the hooked structures known as barbs are present in some types of feathers, and how exactly feathers covered dinosaurian bodies. Göhlich and Chiappe propose that *Juravenator* could be evidence that the bodies of some basal coelurosaurs were more scaly than those of living birds, which have scales only on their lower legs. This explanation is the most likely, given that the authors have carefully examined and analysed the specimen. Yet other possibilities exist.

For example, if *Juravenator* in fact lies more basally than the Compsognathidae in the theropod tree (Fig. 1), we would have a clearer picture of feather evolution instead of this complicated story. We know that feathers appeared

at least as early as the divergence of the Compsognathidae from the main trunk of the theropod tree, but we don't know yet when they first evolved, owing to the poor fossil record. The scaled *Juravenator* would then be a starting point for feather evolution.

In my opinion, this possibility cannot yet be ruled out. First, the evidence² supporting a compsognathid affinity for *Juravenator* is not very strong. Second, the description of *Juravenator* is based on a juvenile specimen, some features of which are likely to have pulled the species up the tree if *Juravenator* followed the developmental patterns revealed by certain other theropod specimens¹². Finally, although they agree in most respects, the proposed relationships^{6–9} between the theropods themselves are not entirely stable.

Whatever the explanation, our knowledge of early feather evolution has been enriched by the discovery of *Juravenator*. Moreover, the discovery emphasizes where future research might fruitfully be concentrated. The Middle-to-Late Jurassic, between about 176 million and 146 million years ago, was a critical time for the origins and early evolution of the coelurosaurian lineages, including birds. But the fossil record is poor, and any new coelurosaur from this period adds considerably to our understanding of the group. *Juravenator* may complicate the picture, but it makes it more complete and realistic.

Xing Xu is at the Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, 142 Xiwai Street, Beijing 100044, China. e-mail: xu.xing@ivpp.ac.cn

- Norell, M. & Xu, X. *Annu. Rev. Earth Planet. Sci.* **33**, 277–299 (2005).
- Göhlich, U. B. & Chiappe, L. M. *Nature* **440**, 329–332 (2006).
- Prum, R. O. & Brush, A. H. *Q. Rev. Biol.* **77**, 261–295 (2002).
- Xu, X. *Integrative Zool.* (in the press).
- Chen, P. J., Dong, Z. M. & Zhen, S. N. *Nature* **391**, 147–152 (1998).
- Sereno, P. C. *Science* **284**, 2137–2147 (1999).
- Norell, M. A., Clark, J. M. & Makovicky, P. J. in *New Perspectives on the Origin and Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 49–67 (Yale Univ. Press, 2001).
- Holtz, T. R. Jr *Geol.* **15**, 5–61 (2000).
- Rauhut, O. W. M. *The Interrelationships and Evolution of Basal Theropod Dinosaurs Spec. Pap. Palaeontol.* **69** (Palaeont. Ass., London, 2003).
- Yu, M.-K., Wu, P., Widelitz, R. B. & Chuong, C.-M. *Nature* **420**, 308–312 (2002).
- Wellnhofer, P. in *Feathered Dragons — Studies on the Transition from Dinosaurs to Birds* (eds Currie, P. J., Koppelhus, E. B., Shugar, M. A. & Wright, J. L.) 282–300 (Indiana Univ. Press, 2004).
- Xu, X. et al. *Nature* **439**, 715–718 (2006).

ANALYTICAL CHEMISTRY

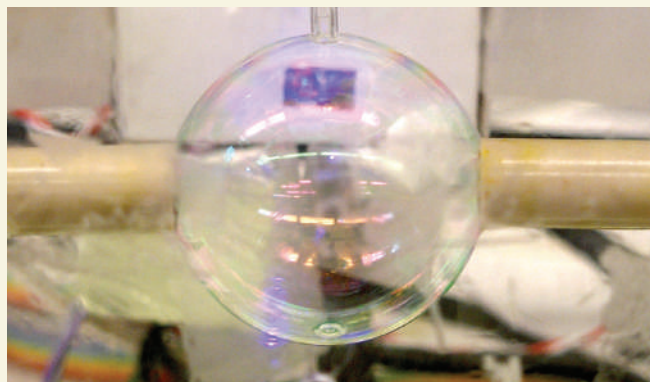
Forever blowing bubbles

The delightful iridescence and evanescence of soap bubbles arises from the fact that their walls are nothing but a film of water sandwiched between two layers of surfactant molecules. This unusual construction also endows soap bubbles with a uniquely large ratio of surface area to contained liquid. Tinakorn Kanyanee *et al.* (*Anal. Chem.* doi:10.1021/ac052198h; 2006) set out to exploit this feature — by ‘wiring up’ soap bubbles and using them as the functional heart of a fully automated detection system for trace gases.

The authors’ bubble factory uses a sealed, transparent plastic box with water at the bottom to keep the environment moist and so prolong bubble life. Soap solution is first supplied to the box through the inner of two concentric tubes. As the

solution spreads from this tube’s slightly recessed tip, it coats the tip of the surrounding outer tube, through which a controlled blast of air is delivered, inflating the soap film and creating a bubble. Precise metering of the air pulse ensures that the bubble is inflated just enough to touch two opposing stainless-steel electrode rods that jut into the box. The bubble-box reproducibly generates long-lived bubbles of uniform size and wall-thickness, so that the conductance measured between the electrodes for a given bubble geometry depends on the bubble’s chemical composition.

To use their system for trace-gas analysis, the authors added the oxidizing agent hydrogen peroxide to the soap solution and exposed the resulting bubble to air containing



traces of sulphur dioxide. The sulphur dioxide readily diffused into the bubble, where it was oxidized and sulphuric acid was produced. The presence of acid increased the bubble’s electrical conductance.

This effect enabled quantitative determination of part-per-billion levels of sulphur dioxide within minutes. The approach should also be applicable to a range of other soluble and reactive gases. But as

Kanyanee and colleagues point out, the wider message is that soap bubbles are an attractive, yet largely unexplored, tool for concentrating trace gases to enable their detection. Whether these species are then detected through conductance measurements, or through bursting the bubble and analysing its liquid by other means, there is, in that sense, more to this story than mostly froth and bubble.

Magdalena Helmer

T. KANYANEE

QUANTUM PHYSICS

A ménage à trois laid bare

Brett D. Esry and Chris H. Greene

Quantum bodies that can’t settle down together in pairs get on fine in a cosy threesome. This startling claim about the private life of particles has just seen its first experimental confirmation.

In 1970, Vitaly Efimov, possessor of a freshly minted Russian PhD in theoretical nuclear physics, predicted a bizarre quantum-mechanical effect¹: a system consisting of three particles, none of whose two-particle subsystems is stable, can, under certain circumstances, produce an infinite number of bound energy levels. That prediction has been a source of concern for theorists ever since. Yet time and again, attempts to disprove the ‘Efimov effect’ have only verified its existence, albeit — until now — purely theoretically. In this issue, more than 35 years on from the initial prediction, Kraemer *et al.* (page 315)² present the first convincing experimental evidence for the effect.

The question explored by Efimov was simple: if each possible pair of particles in a three-particle system is just shy of binding (so there are no bound states), what is the nature of the energy spectrum of the three particles? The answer, that the number of bound states in which the three particles can exist is infinite, is initially surprising. But it can be rationalized by noting that if two particles are already infi-

tesimally close to forming a bound state, just the faintest whiff of an additional attraction will be sufficient to push them over the edge to confederation. A third attracting particle accomplishes precisely that, no matter how far away from the other two it may roam, or how weak its additional attraction may be. Unsurprisingly, given this qualitative argument, three-body Efimov states have been theoretically found³ to be extremely ‘floppy’, with all imaginable triangular shapes — equilateral, isosceles, scalene (Fig. 1, overleaf) — and even linear configurations being comparably probable.

The Efimov effect is distantly related to Llewellyn Hillel Thomas’s 1935 proof⁴ that, in its ground state, a system of three particles that interact only within a vanishingly small range collapses to an infinitely small size, and its binding energy becomes infinitely large. Efimov discussed the link particularly clearly in a 1971 article⁵ that derived an effective potential-energy curve for such a system as a function of a three-particle ‘hyperradius’, R , which is proportional to the root-mean-square distance of the three particles from their centre of mass.

The three-body potential-energy function turns out to be proportional to R^{-2} , and it has a universal, negative coefficient of proportionality. The properties of such a potential are well known, because a similar form governs the motion of a charged particle in the field of a permanent dipole (where two charges of equal and opposite sign are separated by a small distance).

The allowed energy levels of a quantum system can be determined by solving the radial, time-independent Schrödinger equation (which yields the energy levels, E_n , of a quantum system). When the equation is solved for a system with an R^{-2} potential, the allowed energy levels that result follow on iteratively from each other according to the rule $E_{n+1} = E_n \exp(-2\pi/s_0)$, where s_0 is a constant related to the strength of the so-called effective dipole moment with a value slightly greater than one. Thus, as n increases, the binding energies of successive levels decrease exponentially. This scaling also carries over to other observable parameters, such as the size of the state. Thomas’s collapsed states are the infinitely tightly bound states that occur as n approaches $-\infty$, whereas the Efimov states are those states that are only just bound, arising as n approaches $+\infty$. In real physical systems, the fact that the range of interparticle interactions can never be zero prevents a Thomas collapse, so this effect will, one presumes, never be observed. There is, however, no similar obstacle to observing the Efimov effect.

The strength of the interparticle interaction is described in terms of a parameter known as the scattering length, a . In a system consisting of two atoms, if a is large and positive, a weakly bound state of the diatomic molecule exists. If a is negative, the atoms experience an attraction but are not quite bound together. Kraemer and colleagues' experimental work² relied on manipulating the atom–atom scattering length of caesium atoms by the now almost routine technique of tuning a magnetic field to a value very close to a 'Feshbach resonance' — a short-lived, low-energy, two-body molecular state⁶. Controlling the scattering length is crucial, as the formation of Efimov states is only expected for scattering lengths much greater in magnitude than the range of two-body interactions.

Kraemer and colleagues did not actually measure bound states at all, so the simplest 'smoking gun' evidence for Efimov states — a pattern of energy levels that fits the characteristic exponential formula — has yet to be observed. The authors studied instead the low-energy interactions of three ultracold caesium atoms, measuring the three-body recombination process $\text{Cs} + \text{Cs} + \text{Cs} \rightarrow \text{Cs}_2 + \text{Cs}$, which had already been tackled theoretically^{7–10}. Crucially, in 1999, Efimov states were predicted⁸ to show up as an infinite series of resonances in the rate of recombination at negative values of the scattering length, a link confirmed independently¹⁰ two years later.

Kraemer *et al.*² saw only one such resonance (Fig. 2), at a scattering length of $-850a_0$, where $a_0 = 0.53 \text{ \AA}$ is the Bohr radius, a standard unit of distance in atomic physics. The fact that they did not see more (according to the theory, the next resonance should appear at a scattering length 22.7 times more negative, near a scattering length of $-19,000a_0$) leaves them open to

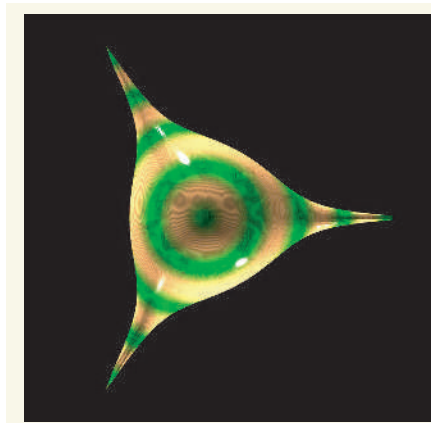


Figure 1 | Floppy triangles. The Efimov-state probability density as a function of two 'hyperangles' that control the shape of the three-particle triangle. In this representation, the radius in any direction from the centre of the object measures the probability density of a particular triangular shape; the spikes indicate a somewhat higher probability for the system to exist in configurations where two of the three atoms are close to each other, whereas the near-spherical part of this object shows that all other triangle shapes occur with very comparable probabilities.

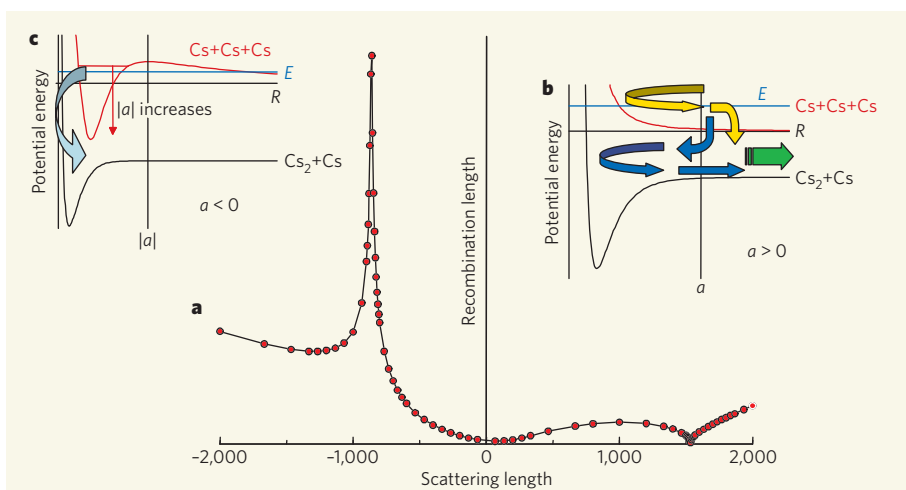


Figure 2 | Features predicted^{7–11} and observed² in ultracold three-body recombination. **a**, The theoretical rate (in terms of a 'recombination length'⁸) of the three-body recombination process $\text{Cs} + \text{Cs} + \text{Cs} \rightarrow \text{Cs}_2 + \text{Cs}$, obtained from numerical solutions of the three-body Schrödinger equation^{8,11} and plotted as a function of the two-body scattering length a (in units of the Bohr radius a_0). The first interference minimum^{7–11} at $a > 0$ and the first resonance maximum^{8–10} at $a < 0$ — features predicted by theory — were observed by Kraemer *et al.*² in their studies of caesium recombination. Insets **b,c**, the three-body potentials responsible for these features: the potential before recombination of three incident atoms at energy E (blue line) is shown in red; that after recombination, in black. **b**, For positive a , the transition between the two potentials takes place at a 'hyperradius' $R \sim a$ by one of two paths. First, the system 'jumps' into the lower recombined channel when still decreasing in size to $R \sim a$. The atom then rebounds off the dimer (R increases) but the system remains in the recombined state (blue pathway). Second, the three atoms initially rebound elastically and only recombine in their outward traversal of $R \sim a$ (yellow path). Minima and maxima in the recombination rate arise from quantum-mechanical interference between these two paths in the outgoing channel (green arrow). **c**, For $a < 0$, the transition occurs at R much smaller than $|a|$, and the system must tunnel quantum mechanically into the small- R potential well in order to recombine. When there are resonances (an allowed quasi-bound energy state, indicated by the horizontal red line) behind this barrier, the tunnelling probability is effectively enhanced, giving a maximum in the recombination rate. The resonance positions change with the scattering length (shown by the arrow), tuning the system in and out of resonance and yielding a series of peaks in the recombination length.

the charge that the effect could be some unrelated, non-Efimov three-body resonance effect, or even just a two-body effect.

That interpretation is rendered unlikely, however, by the appearance of an interference minimum (Fig. 2) in the experimental spectrum at a positive scattering length, $a^{\text{min}} = 210a_0$. Theoretical studies^{7–10} agree that, when Efimov physics occurs, a series of such minima in the three-body recombination should appear for large positive values of a . However, one study¹⁰ places the first minimum correctly to within 25% of the experimentally observed location², even though its formula would not be expected to be valid at such low values of a . On the other hand, numerical solutions of the three-body Schrödinger equation^{8,11} place this minimum at a different location, near $a^{\text{min}} = 1,500a_0$ (Fig. 2). Resolution of this disagreement will require further work.

Although the evidence is not yet conclusive, the appearance of both predicted phenomena — a resonance maximum^{8,10,11} and an interference minimum^{7–11} — implies strongly that Efimov physics has been observed. Developments in the theory are pointing the way to further experiments that would provide even more definitive confirmation¹¹.

The Efimov effect, so long eluding observation, is beginning to yield up its secrets. As our theoretical understanding of Efimov states

has deepened, so has our vision of their wider implications, such as those for the recombination physics studied by Kraemer *et al.*². We look forward to seeing how the spark provided by this impressive new experiment ignites a new round of exploration into the rich quantum-mechanical world of exotic few-body systems.

Brett D. Esry is in the Department of Physics, Kansas State University, Manhattan, Kansas 66506, USA.

e-mail: esry@phys.ksu.edu

Chris H. Greene is in the Department of Physics and JILA, University of Colorado, Boulder, Colorado 80309-0440, USA.

e-mail: chris.greene@colorado.edu

1. Efimov, V. *Phys. Lett. B* **33**, 563–564 (1970).
2. Kraemer, T. *et al.* *Nature* **440**, 315–318 (2006).
3. Esry, B. D., Lin, C. D. & Greene, C. H. *Phys. Rev. A* **54**, 394–401 (1996).
4. Thomas, L. H. *Phys. Rev.* **47**, 903–909 (1935).
5. Efimov, V. N. *Sov. J. Nucl. Phys.* **12**, 589–595 (1971).
6. Inouye, S. *et al.* *Nature* **392**, 151–154 (1998).
7. Nielsen, E. & Macek, J. H. *Phys. Rev. Lett.* **83**, 1566–1569 (1999).
8. Esry, B. D., Greene, C. H. & Burke, J. P. Jr *Phys. Rev. Lett.* **83**, 1751–1754 (1999).
9. Bedaque, P. F., Braaten, E. & Hammer, H.-W. *Phys. Rev. Lett.* **85**, 908–911 (2000).
10. Braaten, E. & Hammer, H.-W. *Phys. Rev. Lett.* **87**, 160407 (2001).
11. D'Incao, J. P. & Esry, B. D. *Phys. Rev. A* (in the press); preprint at www.arxiv.org/cond-mat/0508474 (2006).

OBITUARY

Pierre Potier (1934–2006)

Pharmacist and natural-products chemist, who devised widely used treatments for cancer.

Pierre Potier, who died on 3 February 2006 at the age of 71, was at the forefront of the French school of natural-products chemistry during the second half of the twentieth century. He was a leading light in the development of anticancer drugs from natural sources, and also in encouraging communication between the public and private sectors. In France, these circles tend not to mix. Potier showed that cooperation between the CNRS, the main agency of publicly funded research, and industry could be highly fruitful if the two communities shared their knowledge and experience.

The son of a watchmaker, Potier was born in Bois-Colombes, a northern suburb of Paris. He first trained as a pharmacist, graduating in 1957 from the Faculty of Pharmacy at the University of Paris. He then completed his PhD in chemistry as a student of Maurice-Marie Janot and Jean le Men, taking his first steps in the field of drug synthesis from natural products under the tutelage of Janot at the CNRS's new Institute of Natural Products Chemistry in Gif-sur-Yvette. Following Janot's death, Potier took up the baton in collaborating with the likes of Edgar Lederer, Derek Barton and Guy Ourisson on various projects. Potier himself first co-headed, then headed the institute from 1974 to 2000, and so was responsible for guiding the research of several hundred investigators in learning from, and imitating, nature.

Potier was himself a highly innovative scientist. In 1965, he devised a modification of the Polonovski reaction, which is used to obtain the *N*-demethylation (or *N*-deacylation) of various plant alkaloid derivatives. The modification consisted of replacing one reagent (acetic anhydride) with another (trifluoroacetic anhydride), which facilitated the biomimetic synthesis of many indolic natural products that could not otherwise be made. This reaction was central to the later work involving the 'vinca' alkaloids.

In 1976, Potier and colleagues created a new tool for drug development: the tubulin test for identifying anticancer treatments. Tubulin is the main component of microtubules, which have a central role in cell division. The test was devised by Potier and Daniel Guénard. It does not require sophisticated lab conditions, and allows the swift and cheap evaluation of a drug's capacity to inhibit cell division — and thus of its potential anticancer activity.

The tubulin test led Potier's team to

identify the anticancer activity of vinorelbine (Navelbine), a semi-synthetic derivative of the vinca alkaloids vincristine and vinblastine found in periwinkles. All three drugs inhibit cell division, and have been part of anticancer treatments since the mid-1970s. The production of vinorelbine was taken up by the Pierre Fabre Laboratories for use in treating lung cancer and breast cancer in particular. (Poignantly, Potier's first wife Marie-France, herself a pharmacist, had died of breast cancer in 1968. Potier remarried, and is survived by his second wife Odette and three children.)

But perhaps Potier's greatest contributions came between 1979 and 1994 from a succession of discoveries involving the chemistry of the yew tree. With Guénard and Françoise Guéritte, he first synthesized docetaxel (Taxotere) using a semi-synthetic pathway based on a reaction with 10-deacetylbaccatin III extracted from the needles of the European yew. Taxotere, which was developed with the Rhône-Poulenc Rorer laboratories, is a highly potent antineoplastic drug used mainly against breast, prostate and lung cancer. With Andrew Greene, and using the same chemical steps, the same team then found a way to produce large amounts of paclitaxel (Taxol). The drug had previously been made from the bark of Pacific yew, so that entire trees had to be sacrificed.

More recently, Potier became involved in studying the origin of the degenerative effects of diabetes. He identified a role for glyoxal and methylglyoxal (end-products of carbohydrate metabolism) in causing the enhanced protein degradation seen in diabetic patients. These molecules damage proteins unless they are detoxified by mechanisms involving glyoxalases. Potier patented several drug candidates for treating diabetes in 2005, but at the time of his death was still carrying out experimental investigations on them.

Potier was an author on more than 400 publications and several dozen patents. Ever keen to encourage international cooperation, he created the Franco-Japanese Society of Fine and Medicinal Chemistry, and the Franco-American Chemistry Society. Both nationally and on the global stage, he was especially effective in fostering the collaboration between chemists, biologists and clinicians that is so necessary for successful drug development, and in encouraging links



C. MARMONTEIL

across the academic-industrial divide. It was this background that perhaps led to Potier's appointment by the French government as director-general for research and technology at the Ministry of Research. He held this position from 1994 to 1996, giving priority to promoting publicly funded research and the status of science in France.

He retired from the Institute of Natural Products Chemistry five years ago, but remained emeritus director. On his retirement, he noted with justifiable pride that the royalties paid by pharmaceutical companies marketing the drugs he discovered were able to fund the entire institute.

Pierre Potier became an eminent man, with connections among top politicians and scientists in France and internationally. He received many prizes, among them the CNRS gold medal and the Ernest Guenther Award in the chemistry of natural products; he was also a member of the French Academy of Sciences and the National Academy of Pharmacy, as well as of five foreign academies. Yet he didn't take himself too seriously, retaining the gifts of whimsy and a readiness to laugh. Claude Bernard, a towering figure in physiology during the nineteenth century, was among the first Frenchmen to dream of reconciling the aims and methods of chemists and biologists. With his acute scientific brain and cross-cultural awareness, allied with his sense of humour, generosity and deep appreciation of literature and history, Potier holds a prominent place among Bernard's successors.

François Chast

François Chast is in the Department of Pharmacy, Pharmacology and Toxicology, Hôtel-Dieu Hospital, 1 place du Parvis Notre-Dame, 75004 Paris, France.
e-mail: francois.chast@htd.aphp.fr

NEWS & VIEWS FEATURE

DINOSAUR LOCOMOTION

Beyond the bones

John R. Hutchinson and Stephen M. Gatesy

How did dinosaurs stand and move? Computer simulation and other methods have told us much about how dinosaurs did and did not move, but they have not yet reached their full potential.

Ray Harryhausen, famous animator of creatures in such films as *Jason and the Argonauts* and *The Valley of Gwangi*, has written that “I found dinosaurs to be the ideal subject for stop-motion animation... The fact that nobody knew how those huge reptiles had moved or, for that matter, exactly how they looked meant that I could bring them alive without any fear of criticism”¹. Harryhausen indeed showed that nothing brings an extinct animal to life more than seeing it move. But can walking and running dinosaurs be animated with scientific rigour rather than just artistic flair? Dinosaurs are excellent study organisms for researchers interested in such aspects of biomechanics as the repeated evolution of large size, shifts to and from bipedalism and quadrupedalism^{2,3}, and changes in joint structure, musculature and behaviour^{4–10}.

Unlike investigators working exclusively on living animals, palaeobiologists can’t directly record the motion, measure the forces or dissect the soft-tissue anatomy of their study organisms. Rather than consigning this problem to the unknowable, we can make general inferences about individual species^{11–15}, as well as about large-scale patterns of locomotor evolution^{3,5,10,16}. But specific hypotheses about dinosaur motion, mechanics and performance (such as speed, acceleration and agility) remain controversial^{17,11,13,15,17,18}. Several approaches to the generation and testing of such hypotheses are described in Box 1. Here we focus on a recent trend in the field: the increased use of computer animation and simulation to recreate dinosaur locomotion, taking *Tyrannosaurus rex* as our example.

Uncertainties

Uncertainty about movement stems directly from the framework of the animal’s leg. Bony articulations (Fig. 1a) provide crucial evidence, but deciding which subset of potential joint positions a dinosaur habitually used to stand or move is by no means straightforward. Redundancy in the limb skeleton (what engineers call excess degrees of freedom) precludes a unique solution. For example, how high was the hip joint above the ground (Fig. 1b)? The bones alone do not establish this seemingly

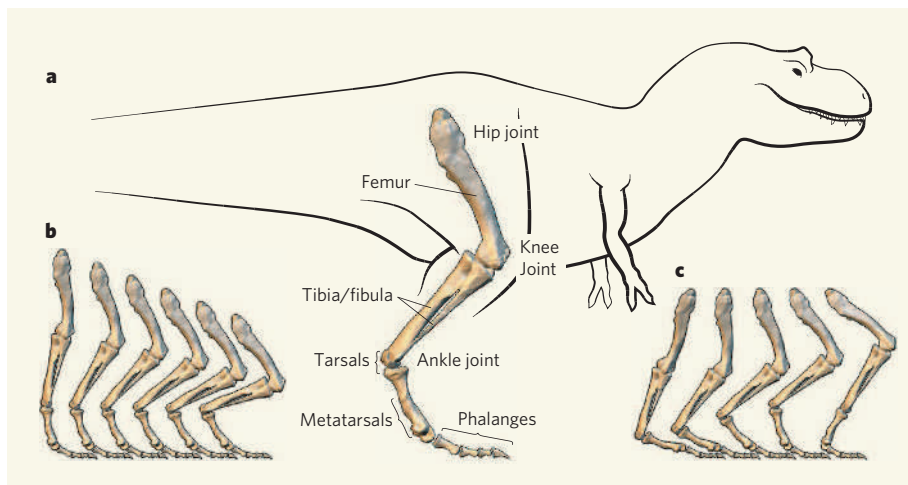


Figure 1 | *Tyrannosaurus rex* as an example of dinosaur anatomy and locomotion. a, The basic skeletal components and joints of the hindlimb. Only the third toe is shown for clarity. **b**, Redundancy allows the limb to assume a range of potential hip heights. **c**, For each position of the hip with respect to the foot, a spectrum of internal configurations is possible.

simple metric, because the joints are flexible enough to allow the tyrannosaur to lie down. Even after a hip position is chosen, a spectrum of poses still remains, because three segments, rather than one or two, span from hip to toes (Fig. 1c). If a dinosaur’s hip, knee, ankle and main toe joints can flex and extend through an arc of 90°, variation of each by 1° increments would yield more than 67 million possible poses. Most of these can be excluded as geometrically or biologically implausible. Yet even in this two-dimensional example, thousands of possible configurations remain.

Animation

The availability of sophisticated animation and simulation software allows palaeobiologists to tackle fundamental questions about dinosaur locomotion in a new way. However, digital dinosaurs are most commonly seen in the popular media, and more often than not little scientific substance underlies such films and documentaries. Yet making virtual dinosaurs move can be done scientifically. This is an excellent opportunity to let the public see science in action.

We had first-hand experience of the complexities of animating a walking *Tyranno-*

saurus as consultants for the ‘Dinosaurs: Ancient Fossils, New Discoveries’ exhibit at the American Museum of Natural History¹⁹. Having previously avoided complete reconstructions in our own research, we wondered whether we could legitimately present our ‘result’ as a scientific hypothesis rather than just a pretty moving picture. We found the process humbling — not because creating a walking stride for the model was difficult, but because (with the aid of the museum’s animator, Scott Harris) inventing motion was all too easy. How would we know which limb poses were off the mark and which were closer to reality? We reluctantly settled on one animation that captured the basic principles of animal locomotion (see Movie 1 in Supplementary Information). Yet we could have produced thousands of animations that were no better or worse. Computational tools offer tremendous power to generate motion, but we now face the new problem of how to choose among countless hypotheses.

Palaeontologists have two main options when confronting redundancy in an articulated chain of bones (Fig. 1). The first is to set arbitrary constraints on the system to reduce the number of possibilities. A dinosaur’s

Box 1 Traditional approaches to investigating dinosaur locomotion⁶

● **Comparison with living analogues** (for example, hoofed mammals, birds, elephants and rhinoceroses) reveals some broad-brush similarities^{4,7}. But there are no rigorous means of testing the choice of analogue and the extent to which its characteristics are applicable. Once analogies are inferred, little further progress is possible.

● **Bone scaling** is a more quantitative method, revealing how dinosaur bone shape changed with size. It has shown that size had the same influence on dinosaurs as on many mammals^{2,6,27}, strengthening the conclusion that larger dinosaurs were not as athletic as smaller ones. Yet this yields little specific information about how any one dinosaur moved.

● **Fossil footprints** are stunning evidence of dinosaur behaviour^{8,25,28,29}, for example supporting the inference that ancestral dinosaurs stood and moved with more erect bipedal limbs than their more sprawling and quadrupedal ancestors³⁰. But estimating speed from footprints has a wide margin of error¹², and footprints tell us little about how the whole limb or body moved, because of redundancy within the limb skeleton (Fig. 1).

● **Functional morphology** is a classical foundation of locomotion research. It is important for determining how much motion or what poses certain limb joints could allow^{4,7,9,30}. Much like scaling approaches, bones are usually

emphasized over soft tissues. Limb-joint redundancy is again a serious problem.

● **Biomechanical modelling** is one of the most rigorous methods to reconstruct how specific dinosaurs moved^{11–15,17,18,20,21}. Locomotor complexity cannot be captured in simple models, however, and the number of unknown parameters jumps quickly as models become more complex. The best models are validated by data from living animals and analyse unknown parameters to see if reasonable values would drastically change the results.

● **Viewing dinosaurs in their evolutionary context** tells us much about broad historical trends but little about how specific animals moved. For example, within the theropod dinosaurs, which include *Tyrannosaurus*, tracing the skeletal correlates of limb muscle attachments reveals a stepwise pattern of change, with the hindlimb attaining its 'modern' condition only in animals closely related to extant birds^{3,5,6,10,28}. Yet how did any one dinosaur fit into this trend? Did *Tyrannosaurus* move its femur through large arcs during walking (like a crocodile), as we presume the ancestors of both of these groups of reptiles did? Or can we identify aspects of its locomotion that were more derived, closer to the condition in extant birds, such as keeping the femur more horizontal during walking, entailing a more crouched limb configuration?

J.R.H. & S.M.G.

are placed on the musculoskeletal system. The GRF increases with speed, and so at top speed the maximum GRF can be used to estimate the size of the muscles needed to support the limb^{13–15}.

Constraints

A range of kinematic (motion-based or geometric) and kinetic (force-based) criteria help to exclude certain poses for *Tyrannosaurus* as unrealistic. The goal is to constrict the 'configuration space' of poses as narrowly as possible, by sequentially adding constraints that rule out poses as impossible or unlikely, depending on the strength of support from studies of the locomotion of extant animals. We add kinematic constraints first. For example, body segments cannot penetrate each other or the surface of the ground, and joints cannot exceed certain maximal and minimal joint angles — in particular, the ankle and knee cannot bend backwards. But a huge configuration space remains, and many plausible poses cannot be excluded (Fig. 3).

Next, kinetic constraints based on biomechanical studies are added. First, the GRF is oriented vertically through its point of application at the foot–ground interface^{24,25}. Second, poses are excluded in which any joints have unreasonable joint moments incurred by the GRF — the hip, knee and ankle require the action of extensor muscles that straighten the limb and work against the force of gravity. Again, this constraint is congruent with living animals¹⁶. It excludes the first example pose (Fig. 3a), because the GRF is in front of the knee, requiring net knee flexor (bending) muscle activity to balance it, not knee extensors. The third kinetic constraint is that the moment arm of the GRF about any joint cannot exceed a maximal distance. This is determined by the maximal extensor muscle moment that could be generated by an animal with 5% of its body mass dedicated to muscles crossing that joint^{14,15}. This maximal distance is inversely related to the magnitude of the GRF and excludes the second pose (Fig. 3b); the dinosaur would not have enough knee extensor muscles even to run slowly.

These realistic kinetic constraints involve more assumptions than the kinematic ones. They exclude many more poses than are shown here; yet many remain, such as Figure 3c. That pose allows running, although not very quickly (no more than 29 km h⁻¹), as the hip extensor muscles would not have been able to sustain higher speeds under the assumptions invoked.

The challenges don't end with identifying candidate mid-stance poses. The remainder of the stance phase and the swing phase expand the potential motion space of the limb enormously. Once such an analysis moves beyond a single pose, the issue of time must be dealt with (see Movie 2 in Supplementary Information). One can calculate the stride length for any motion (Fig. 2a). To estimate the velocity,

limb orientation is often taken for granted by duplicating a mammalian, avian or some other intuitively pleasing pose^{7,20}. Other studies invoke simple geometric rules: for example, the knee and ankle can be restricted to identical angles, or the hips fixed midway between the planted feet²¹. Such constraints simplify the redundancy enough to calculate a narrow range of solutions, but one risks answering a question that no longer has relevance to reality¹⁸. Assuming a pose *a priori* is untenable.

Beyond the bones

The second, alternative option is to use demonstrably realistic constraints to exclude, rather than include, possible limb motions. Only by admitting that a single solution is unachievable can we begin to search for a set of movements that cannot be ruled out. A rigorous dynamic simulation^{22,23} of a moving dinosaur, one encompassing all motions and forces, cannot yet plausibly be done. But a major goal is to move towards an integrative analysis based on solid evidence and an adequate consideration of unknown factors. Because the limb skeleton is so redundant, this can be done by moving beyond the bones to include the forces (muscular, external, gravitational) that actually create limb motion.

A complete locomotor cycle, or stride, consists of a stance phase and a swing phase

(Fig. 2a). In the stance phase, the limb pushes down on the ground to incur an equal but opposite force (the ground reaction force or GRF; Fig. 2b). In early stance, the GRF points up and back, decelerating the body as the limb shortens. At mid-stance, the GRF is vertical and is at its peak value for a stride. Its relative magnitude is a function of the fraction of a stride that each foot is on the ground. In late stance, the GRF points up and forward, reaccelerating the body^{24,25}. Throughout the stance phase, forces from muscles act at a distance from a joint axis (the moment arm²⁶) to produce moments (rotational forces) that prevent the limb from collapsing. Figure 2c–e shows three possible stance phases for a running *Tyrannosaurus*, none of which can be excluded based on the articulations between bones or other common criteria. The first is a very upright 'columnar' limb²¹. The second represents a more crouched 'bird-like' limb⁷. The third uses poses between these two extremes. How could we know which of these is more or less likely?

Rather than tackle the intricacies of a complete stance phase, a single time point can define the problem better — linking multiple instants of locomotion might allow reconstruction of a full stance phase, then a full stride. We begin with a simple example: a *Tyrannosaurus* frozen at mid-stance (Fig. 3). This can act as a 'filter' for excluding improbable poses, as it is a time when high stresses

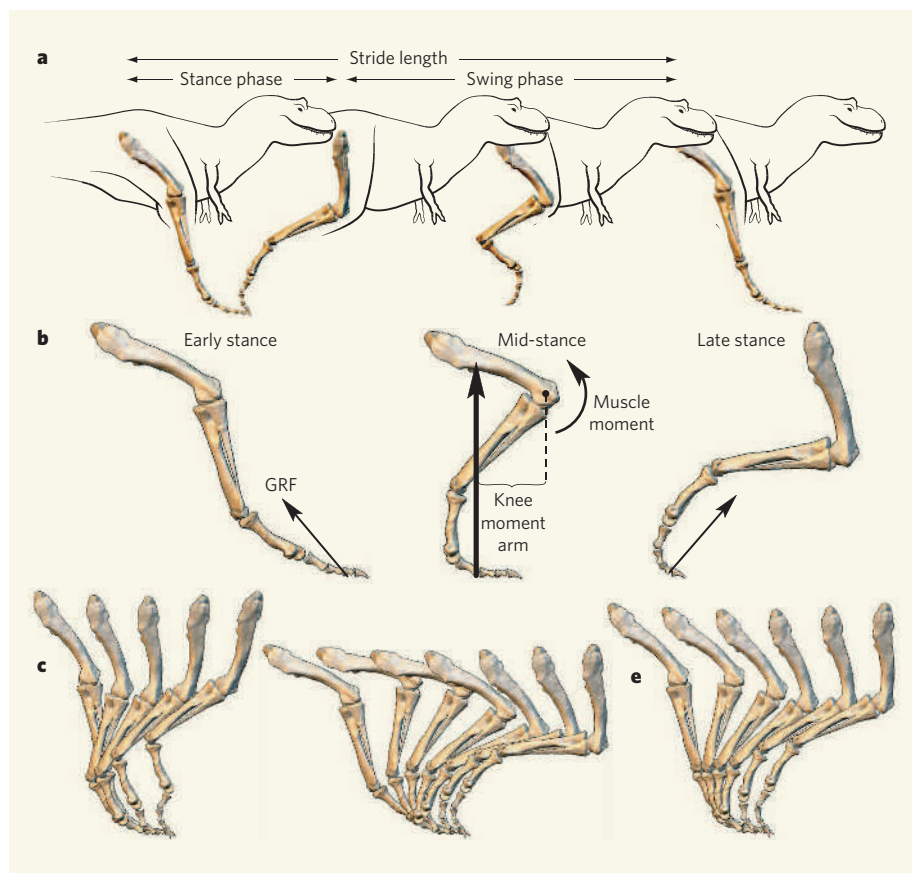


Figure 2 | Locomotor cycle (stride) for *Tyrannosaurus*. **a**, The main components of a stride. **b**, A stance limb induces an equal and opposite ground reaction force (GRF), represented here as a vector. The GRF requires opposing muscle activity to balance it. The horizontal bracket indicates the moment arm of the GRF about the knee joint. **c–e**, Three possible stance-phase movements for a running *Tyrannosaurus*.

however, a time-dependent factor — the stride frequency — is also required, because velocity equals the distance travelled per cycle multiplied by the number of cycles per unit time. How fast do the limbs swing forward through the air, and what are their motions while unconstrained by the GRF? This is one of the most vexing problems for reconstructing

dinosaur locomotion. Potential solutions depend on invoking experimentally validated constraints from living animals, but these constraints are often not well enough understood to apply them to extinct animals. Yet now that researchers know what to look for, better constraints will probably be found.

Back to biology

The subject of dinosaur locomotion requires palaeontologists to become better functional biologists — experts in more than just skeletal morphology. The most appropriate approaches focus on the entire limb rather than just bones. With the tools now available, we can begin to study how fast certain dinosaurs could move, how their muscles functioned, and how they stood or moved. But the problems of missing data (such as muscle size or moment arms^{15,26}) will always remain and cannot be overlooked.

We have provided a case study of how to embrace the uncertainty in reconstructions of the locomotion of extinct animals. The example of mid-stance poses shows that it is possible to move forward without being confounded by the vast reaches of this uncertainty or making arbitrary assumptions, by using a range of kinematic and kinetic constraints that depend on their applicability to living animals.

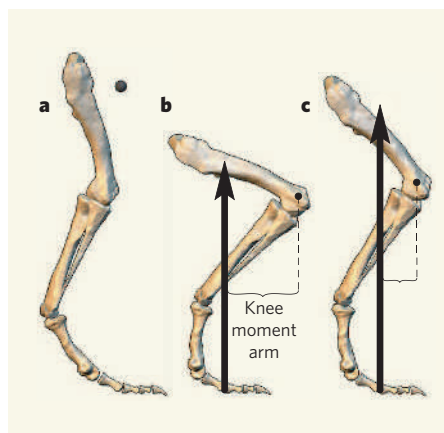


Figure 3 | Mid-stance poses for the three sequences shown in Figure 2. The horizontal bracket indicates the moment arm of the GRF about the knee joint. Kinetic constraints can rule out poses **a** and **b**, and many others, but not pose **c** and many others.

Such an approach to reconstructing behaviour in extinct animals can bolster confidence in the conclusions, and thereby move the field in a more progressive, sustainable direction. In publicly engaging and often fast-moving fields such as dinosaur palaeobiology, it is essential that the questions do not get too far ahead of the evidence, that tides of enthusiasm do not drown concerns about ambiguity, and that novel tools are not recklessly applied. The main principles of this approach apply not only to dinosaur locomotion, but also to all functional studies of extinct organisms — from breathing and feeding, to swimming and flight.

Ray Harryhausen preferred animated dinosaurs to retain “a look of pure fantasy because their movements are beyond anything we know”¹. We may indeed never know for sure how dinosaurs moved. But if we can find out what their movements were not like, that will be real progress.

John R. Hutchinson is in the Structure and Motion Laboratory, Department of Veterinary Basic Sciences, The Royal Veterinary College, London NW1 0TU, UK.

Stephen M. Gatesy is in the Department of Ecology and Evolutionary Biology, Box G, Brown University, Providence, Rhode Island 02912, USA.

e-mail: jrthutch@rvc.ac.uk

1. Harryhausen, R. & Dalton, T. *Ray Harryhausen: An Animated Life* 8, 21 (Billboard, New York, 2004).
2. Carrano, M. T. *J. Zool.* **254**, 41–55 (2001).
3. Carrano, M. T. *Paleobiology* **26**, 489–512 (2000).
4. Coombs, W. P. Jr. *Q. Rev. Biol.* **53**, 393–418 (1978).
5. Gatesy, S. M. *Paleobiology* **16**, 170–186 (1990).
6. Carrano, M. T. *Paleobiology* **24**, 450–469 (1998).
7. Paul, G. S. *Gaia* **15**, 257–270 (1998).
8. Wilson, J. A. & Carrano, M. T. *Paleobiology* **25**, 252–267 (1999).
9. Farlow, J. O. et al. *Am. Zool.* **40**, 640–663 (2000).
10. Hutchinson, J. R. & Gatesy, S. M. *Paleobiology* **26**, 734–751 (2000).
11. Alexander, R. McN. *Zool. J. Linn. Soc.* **83**, 1–25 (1985).
12. Alexander, R. McN. *Dynamics of Dinosaurs and Other Extinct Giants* (Columbia Univ. Press, 1989).
13. Hutchinson, J. R. & Garcia, M. *Nature* **415**, 1018–1021 (2002).
14. Hutchinson, J. R. *J. Morphol.* **262**, 421–440 (2004).
15. Hutchinson, J. R. *J. Morphol.* **262**, 441–460 (2004).
16. Hutchinson, J. R. *C. R. Palevol.* (in the press).
17. Christiansen, P. *Gaia* **15**, 241–255 (1998).
18. Blanco, R. E. & Mazzetta, G. V. *Acta Palaeontol. Polon.* **46**, 193–202 (2001).
19. www.amnh.org/exhibitions/dinosaurs
20. Blanco, R. E. & Jones, W. W. *Proc. R. Soc. Lond. B* **272**, 1769–1773 (2005).
21. Henderson, D. M. *Ichnos* **100**, 99–114 (2003).
22. Anderson, F. C. & Pandey, M. G. *Gait Posture* **17**, 159–169 (2002).
23. Sellers, W. I. et al. *J. Exp. Biol.* **206**, 1127–1136 (1998).
24. Alexander, R. McN. *Principles of Animal Locomotion* (Princeton Univ. Press, 2003).
25. Biewener, A. A. *Animal Locomotion* (Oxford Univ. Press, 2003).
26. Hutchinson, J. R. et al. *Paleobiology* **31**, 676–701 (2005).
27. Christiansen, P. *J. Vert. Paleontol.* **19**, 666–680 (1999).
28. Day, J. J. et al. *Nature* **415**, 494–495 (2002).
29. Gatesy, S. M. et al. *Nature* **399**, 141–144 (1999).
30. Christiansen, P. *Gaia* **14**, 45–75 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com

BRIEF COMMUNICATIONS

Photocatalyst releasing hydrogen from water

Enhancing catalytic performance holds promise for hydrogen production by water splitting in sunlight.

Direct splitting of water using a particulate photocatalyst would be a good way to produce clean and recyclable hydrogen on a large scale¹, and in the past 30 years various photocatalysts have been found that function under visible light^{2–4}. Here we describe an advance in the catalysis of the overall splitting of water under visible light: the new catalyst is a solid solution of gallium and zinc nitrogen oxide^{5,6}, $(\text{Ga}_{1-x}\text{Zn}_x)(\text{N}_{1-x}\text{O}_x)$, modified with nanoparticles of a mixed oxide of rhodium and chromium. The mixture functions as a promising and efficient photocatalyst in promoting the evolution of hydrogen gas.

$(\text{Ga}_{1-x}\text{Zn}_x)(\text{N}_{1-x}\text{O}_x)$ is a yellow powder with an absorption edge at about 510 nm (sample A, see supplementary information). Used alone, it has little photocatalytic activity. We impregnated $(\text{Ga}_{1-x}\text{Zn}_x)(\text{N}_{1-x}\text{O}_x)$ with nanoparticles of rhodium–chromium mixed oxide (for methods, see supplementary information). Scanning and transmission electron microscopy revealed that fine particles (10–20 nm) were distributed uniformly on the $(\text{Ga}_{1-x}\text{Zn}_x)(\text{N}_{1-x}\text{O}_x)$ surface and that the loaded nanoparticles were crystalline, as indicated by the lattice fringes in the transmission electron micrograph image (for details, see supplementary information). Energy-dispersive X-ray analysis confirmed that these nanoparticles contained both rhodium and chromium.

This photocatalyst modification markedly enhances the overall splitting of water driven by visible light. Modification with either chromium or rhodium oxide alone did not appreciably increase this photocatalytic activity.

Figure 1a shows the quantum efficiency of overall water splitting by the photocatalyst as a function of the wavelength of the incident light. The quantum efficiency decreases with increasing wavelength, and the longest wavelength suitable for overall water splitting coincides with the absorption edge of the $(\text{Ga}_{1-x}\text{Zn}_x)(\text{N}_{1-x}\text{O}_x)$ solid solution; this indicates that the reaction proceeds through light absorption by the solid solution. The quantum efficiency of overall water splitting on this catalyst is about 2.5% at 420–440 nm, which is about an order of magnitude higher than the previously reported activity of photocatalysts used in overall water splitting under visible light^{4,6}.

A typical time course for overall water splitting on the impregnated $(\text{Ga}_{1-x}\text{Zn}_x)(\text{N}_{1-x}\text{O}_x)$

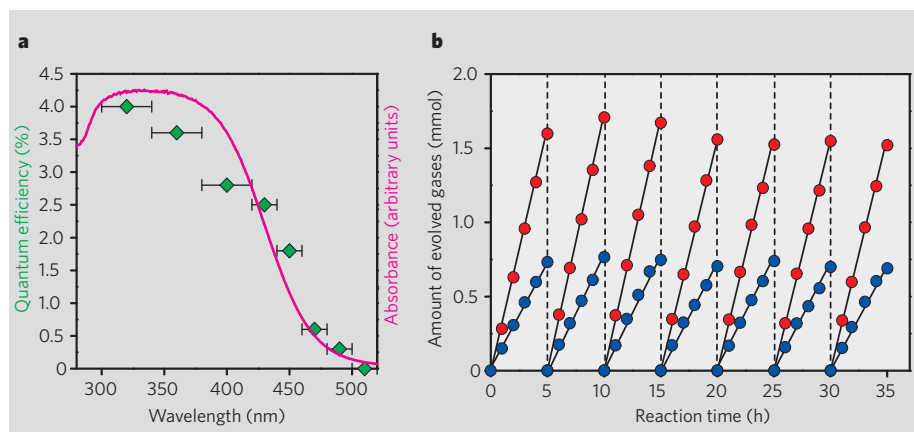


Figure 1 | Photocatalytic performance of the photocatalyst $(\text{Ga}_{1-x}\text{Zn}_x)(\text{N}_{1-x}\text{O}_x)$ loaded with mixed oxides of rhodium and chromium in overall water splitting. **a**, Quantum efficiency of the photocatalyst (sample C, see supplementary information) plotted as a function of wavelength of the incident light. Quantum efficiency was estimated by using several cut-off filters; each plot is in the middle of two cut-off wavelengths (for methods, see supplementary information). **b**, Time course of overall water splitting under visible light (of wavelength longer than 400 nm) on the photocatalyst (sample B, see supplementary information). The reaction was continued for 35 h, with evacuation every 5 h (dotted line). Red circles, hydrogen; blue circles, oxygen.

photocatalyst under visible light (at wavelengths longer than 400 nm) is shown in Fig. 1b. Stoichiometric evolution of hydrogen and oxygen was evident from the start of the reaction, and there was no noticeable degradation of activity in repeated runs for 35 h. The total evolution of hydrogen and oxygen after 35 h was 16.2 mmol, exceeding the amount of catalyst (3.7 mmol) used in the reaction. Radiation from a high-pressure mercury lamp through Pyrex glass of a suspension of the photocatalyst powder in water induced gas evolution as bubbles at atmospheric pressure (for movie, see supplementary information). The photocatalyst is therefore stable during overall water splitting that is driven by visible light.

If silver nitrate is used as a sacrificial electron acceptor, the quantum efficiency for oxygen evolution rises to 51% at 420–440 nm, which is 20 times higher than that for overall water splitting. The photocatalytic activity of the modified $(\text{Ga}_{1-x}\text{Zn}_x)(\text{N}_{1-x}\text{O}_x)$ solid solution could therefore be increased further by improving the hydrogen-evolution site. Its composition could also be modified to extend the absorption edge to longer wavelengths. For now, our results demonstrate the feasibility of using photocatalysts and solar energy to produce hydrogen from water.

Kazuhiko Maeda*, Kentaro Teramura*,

Daling Lu†, Tsuyoshi Takata*, Nobuo Saito‡, Yasunobu Inoue‡, Kazunari Domen*†

*Department of Chemical System Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: domen@chemsys.t.u-tokyo.ac.jp

†Solution Oriented Research for Science and Technology (SORST) programmes of the Japan Science and Technology Corporation, 4-1-8 Honcho, Kawaguchi-shi, Saitama 332-0012, Japan

‡Department of Chemistry, Nagaoka University of Technology, Nagaoka 940-2188, Japan

1. Bard, A. J. & Fox, M. A. *Acc. Chem. Res.* **28**, 141–145 (1995).
2. Fujishima, A. & Honda, K. *Nature* **238**, 37–38 (1972).
3. Kato, H. & Kudo, A. *J. Phys. Chem. B* **106**, 5029–5034 (2002).
4. Sayama, K., Mukasa, K., Abe, R., Abe, Y. & Arakawa, H. *Chem. Commun.* 2416–2417 (2001).
5. Maeda, K. et al. *J. Am. Chem. Soc.* **127**, 8286–8287 (2005).
6. Maeda, K. et al. *J. Phys. Chem. B* **109**, 20504–20510 (2005).

Supplementary information accompanies this communication on Nature's website.

Received 5 December 2005; accepted 20 February 2006.

Competing financial interests: declared none.
doi:10.1038/440295a

BRIEF COMMUNICATIONS ARISING online
www.nature.com/bca see Nature contents.

GEOPHYSICS

Hot fluids or rock in eclogite metamorphism?

Arising from: A. Camacho, J. K. W. Lee, B. J. Hensen & J. Braun *Nature* **435**, 1191–1196 (2005)

The mechanisms by which mafic rocks become converted to denser eclogite in the lower crust and mantle are fundamental to our understanding of subduction, mountain building and the long-term geochemical evolution of Earth. Based on larger-than-expected gradients in argon isotopes, Camacho *et al.*¹ propose a new explanation — co-seismic injection of hot (700 °C) aqueous fluids into much colder (400 °C) crust — for the localized nature of eclogite metamorphism during Caledonian crustal thickening, as recorded in the rocks of Holsnøy in the Bergen arcs, western Norway. We have studied these unusual rocks^{2–4}, which were thoroughly dehydrated under granulite facies conditions during a Neoproterozoic event (about 945 million years (945 Myr) ago); we also concluded that fracture-hosted fluids were essential as catalysts and components in the conversion to eclogite about 425 Myr ago⁵. However, we are sceptical of the assertion by Camacho *et al.* that eclogite temperatures were reached only in the vicinity of fluid-filled fractures. Determining whether these rocks were strong enough to fracture at depths of 50 km because they were cold or because they were very dry is crucial to understanding the mechanics of the lower crust in mountain belts, including, for example, the causes of seismicity in the Indian plate beneath the modern Himalayas⁶.

The interpretation by Camacho *et al.* of their argon-isotope data is inconsistent with reasonable magnitudes of heat and fluid advection, as well as with structural field relations in these rocks, for two reasons. First, immense quantities of fluid — equal at a minimum to the mass of eclogitized rock — would have been required to heat the rocks to 300 °C above the ambient temperature. Moreover, the source of any such superheated fluids is not obvious for the Holsnøy complex. There is no evidence of magmatism at the time of the eclogite metamorphism and, because the rocks resided somewhere within the over-thickened Caledonian crust, any fluids derived from a subducting slab would have had to travel long distances (20 km, assuming a geothermal gradient of 15 °C km⁻¹) without losing their heat in order to trigger eclogite-grade reactions in cold rocks.

Second, Camacho *et al.* suggest that the observed occurrence of eclogite in 10–100-m-wide shear zones could be explained by multiple seismically triggered episodes of hot fluid injection into the same sites. However, consideration of structural relationships make this scenario implausible. Although most of the eclogites occur as shear zones, cross-cutting

relationships show that these shear zones were secondary results of, rather than the primary conduits for, fluid infiltration. Brittle tensile and shear fractures, as well as pseudotachylite (glassy rock generated by frictional melting during seismic slip), are seen exclusively in the almost anhydrous granulite facies rocks, whereas only eclogitized areas show evidence of pervasive ductile strains that overprint the granulite-facies fabric⁵.

These ubiquitous structural relationships indicate that the dry granulite was much stronger than the slightly rehydrated (phengite- and zoisite-bearing) eclogites, which accommodated very large ductile shear strains following their metamorphic conversion. Although it is surprising to see evidence of frictional sliding behaviour at such depths and temperatures, the virtual absence of hydrous phases in the gabbroic to anorthositic granulites can account for their unusual rheological properties at high temperatures^{7,8}.

Once formed, the eclogites on Holsnøy were apparently too weak to fracture, seismically or otherwise, and they responded to far-field stresses instead by ductile deformation — making repeated, rapid infiltration of large volumes of fluid into the same sites unlikely. The resultant shear zones may have been loci of slow, diffuse fluid flow, but under such conditions the fluids would have steadily lost their heat to the surrounding rocks if the ambient temperature had been only 400 °C. Hence, the scenario proposed by Camacho *et al.*¹ cannot readily account for the broadest belts of eclogite,

such as the Hundskjefte shear zone (shown in Fig. 1 of ref. 1), which is as much as 500 m wide. We suspect that the profound decrease in strength upon conversion to eclogite caused the metamorphic process to be self-limiting and that it resulted in the observed 'patchiness' of the eclogitized areas⁵.

We find that the model proposed by Camacho *et al.* is at odds not only with plausible geophysical constraints but also with some of the principal characteristics of this remarkable rock complex, which provides a rare glimpse of how different geological processes may be when rocks are exceptionally dry.

M. G. Bjørnerud*, H. Austrheim†

*Geology Department, Lawrence University, Appleton, Wisconsin 54912, USA
e-mail: bjornerm@lawrence.edu

†Physics of Geological Processes Project, Department of Earth Sciences, PO 1047, University of Oslo, 0316 Oslo, Norway

1. Camacho, A., Lee, J. K. W., Hensen, B. J. & Braun, J. *Nature* **435**, 1191–1196 (2005)
2. Austrheim, H. & Griffin, W. *Chem. Geol.* **50**, 267–281 (1985).
3. Jamtveit, B., Austrheim, H. & Mørth-Sørensen, A. *Nature* **408**, 75–78 (2000).
4. Austrheim, H. & Boundy, T. *Science* **265**, 82–83 (1994).
5. Bjørnerud, M., Austrheim, H. & Lund, M. *J. Geophys. Res.* **107**, 2252–2269 (2002).
6. Jackson, J. A., Austrheim, H., McKenzie, D. & Priestly, K. *Geology* **32**, 625–628 (2004).
7. Mackwell, S., Zimmerman, M. & Kohlstedt, D. *J. Geophys. Res.* **103**, 975–984 (1998).
8. Rubey, D. in *Deformation Processes in Minerals, Ceramics, and Rocks* (eds Barber, D. & Meredith, P.) 262–295 (Chapman and Hall, New York, 1990).

doi:10.1038/nature04714

GEOPHYSICS

Camacho et al. reply

Replying to: M. G. Bjørnerud & H. Austrheim *Nature* **440**, doi:10.1038/nature04714 (2006)

Bjørnerud and Austrheim¹ interpret the geological evidence in the rocks of Holsnøy at Lindås nappe, Norway, to be inconsistent with our cold-crust model², but do not question our new argon isotopic data, on which we base the thermal history of the terrain. A critical flaw underlying their arguments^{1,3} is the implicit assumption that element diffusion does not occur in dry environments, although there is clear evidence to the contrary^{2,4,5}. Counter to earlier claims^{3,6} of element and isotope immobility in dry rocks, we have demonstrated the existence of diffusion profiles in phlogopite associated with the uptake of argon during the Caledonian in 'unreacted' protolith of the

Lindås nappe. Diffusion has taken place in these dry rocks and cannot be ignored.

Addressing their specific objections to our model in turn, we consider first the transport of hot fluids through the crust. Our contention is not that the whole terrain was heated by 300 °C, but rather that regional ambient temperatures were not significantly affected by localized advection of hot fluids. Regarding the source of fluids, it is widely accepted that ample volumes can be provided from dehydration of the subducted rocks and there is nothing to prevent these fluids travelling long distances.

Second, regarding the temporal relationship of fluid access, recrystallization and

deformation, Bjørnerud and Austrheim¹ suggest that the hydrated eclogite formed before ductile deformation (also at eclogite facies). This implies that the hydrated assemblage was overprinted by a later fluid-absent, high-strain event, confined only to the eclogite, for which there is no field evidence. We believe that the structural and microstructural relationships indicate that deformation and fluid infiltration were coeval. Channelized fluid flow through shear zones⁷ accompanying fluid-enhanced dynamic recrystallization⁸ is a widely accepted process.

In further support of our contention that the plagioclase-rich granulites remained cool and were not at 700 °C, as suggested by Bjørnerud and Austrheim^{1,3}, we note that plagioclase deforms plastically at temperatures above about 550 °C (ref. 9): we do not observe this deformation. In fact, the experimental data¹⁰ quoted by Bjørnerud and Austrheim¹ show that, under dry conditions, plagioclase-poor rocks were stronger than plagioclase-rich rocks, with the rock strength defined by dislocation creep in plagioclase.

Contrary to the assertion of Bjørnerud and Austrheim¹, we do not find the occurrence of

frictional sliding at great depths to be surprising, as many deep earthquakes have been recorded in subducting slabs and in the mantle. We can reconcile the geological evidence with recent interpretations of seismic data from seismogenic zones¹¹. After an earthquake of magnitude 8 in Chile in 1995, 4,426 aftershocks were recorded over 3 months and interpreted as representing the rapid migration of fluid into the overlying plate. This is analogous to what could have been occurring in the Bergen arcs during the Caledonian, with large earthquakes producing pseudotachylite and aftershocks resulting from fluid infiltration.

We therefore find that the arguments put forward by Bjørnerud and Austrheim¹ do not provide compelling evidence against our model, nor do they provide an alternative scenario that reconciles all the geological evidence. Our model coherently and consistently integrates the geological observations with isotopic data, geophysical constraints and the tectonic setting of the Bergen arcs.

Alfredo Camacho*, James K. W. Lee*, Bastiaan J. Hensen†, Jean Braun‡

*Geological Sciences and Geological Engineering,

Queen's University, Kingston, Ontario K7L 3N6, Canada

e-mail: camacho@geol.queensu.ca

†School of Biology, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia

‡Géosciences Rennes, CNRS UMR 6118, Université de Rennes 1, Rennes, CS 74205, France

1. Bjørnerud, M. & Austrheim, H. *Nature* **440**, doi:10.1038/nature04714 (2006).
2. Camacho, A., Lee, J. K. W., Hensen, B. J. & Braun, J. *Nature* **435**, 1191–1196 (2005).
3. Bjørnerud, M. & Austrheim, H. *J. Struct. Geol.* **24**, 1537–1538 (2002).
4. Camacho, A., McDougall, I., Armstrong, R. & Braun, J. *J. Struct. Geol.* **23**, 1007–1013 (2001).
5. Camacho, A. *J. Struct. Geol.* **24**, 1539–1540 (2002).
6. Kühn, A., Glodny, J., Iden, K. & Austrheim, H. *Lithos* **51**, 305–330 (2000).
7. Cox, S. F. *Earth Planets Space* **54**, 1121–1125 (2002).
8. Carter, K. E. & Dworkin, S. I. *Geology* **18**, 720–723 (1990).
9. Tullis, J., Stünitz, H., Teyssier, C. & Heilbronner, R. *Stress, Strain and Structure, a Volume in Honour of W.D. Means* (eds Jessell, M. W. & Urai, J. L.) (2000).
10. Mackwell, S., Zimmerman, M. & Kohlstedt, D. J. *Geophys. Res.* **103**, 975–984 (1998).
11. Husen, S. & Kissling, E. *Geology* **29**, 847–850 (2001).

doi:10.1038/nature04715

Folding DNA to create nanoscale shapes and patterns

Paul W. K. Rothemund¹

'Bottom-up fabrication', which exploits the intrinsic properties of atoms and molecules to direct their self-organization, is widely used to make relatively simple nanostructures. A key goal for this approach is to create nanostructures of high complexity, matching that routinely achieved by 'top-down' methods. The self-assembly of DNA molecules provides an attractive route towards this goal. Here I describe a simple method for folding long, single-stranded DNA molecules into arbitrary two-dimensional shapes. The design for a desired shape is made by raster-filling the shape with a 7-kilobase single-stranded scaffold and by choosing over 200 short oligonucleotide 'staple strands' to hold the scaffold in place. Once synthesized and mixed, the staple and scaffold strands self-assemble in a single step. The resulting DNA structures are roughly 100 nm in diameter and approximate desired shapes such as squares, disks and five-pointed stars with a spatial resolution of 6 nm. Because each oligonucleotide can serve as a 6-nm pixel, the structures can be programmed to bear complex patterns such as words and images on their surfaces. Finally, individual DNA structures can be programmed to form larger assemblies, including extended periodic lattices and a hexamer of triangles (which constitutes a 30-megadalton molecular complex).

In 1959, Richard Feynman put forward the challenge of writing the *Encyclopaedia Britannica* on the head of a pin¹, a task which he calculated would require the use of dots 8 nm in size. Scanning probe techniques have essentially answered this challenge: atomic force microscopy² (AFM) and scanning tunnelling microscopy^{3,4} (STM) allow us to manipulate individual atoms. But these techniques create patterns serially (one line or one pixel at a time) and tend to require ultrahigh vacuum or cryogenic temperatures. As a result, methods based on self-assembly are considered as promising alternatives that offer inexpensive, parallel synthesis of nanostructures under mild conditions⁵. Indeed, the power of these methods has been demonstrated in systems based on components ranging from porphyrins⁶ to whole viral particles⁷. However, the ability of such systems to yield structures of high complexity remains to be demonstrated. In particular, the difficulty of engineering diverse yet specific binding interactions means that most self-assembled structures contain just a few unique positions that may be addressed as 'pixels'.

Nucleic acids can help overcome this problem: the exquisite specificity of Watson–Crick base pairing allows a combinatorially large set of nucleotide sequences to be used when designing binding interactions. The field of 'DNA nanotechnology'^{8,9} has exploited this property to create a number of more complex nanostructures, including two-dimensional arrays with 8–16 unique positions and less than 20 nm spacing^{10,11}, as well as three-dimensional shapes such as a cube¹² and truncated octahedron¹³. However, because the synthesis of such nanostructures involves interactions between a large number of short oligonucleotides, the yield of complete structures is highly sensitive to stoichiometry (the relative ratios of strands). The synthesis of relatively complex structures was thus thought to require multiple reaction steps and purifications, with the ultimate complexity of DNA nanostructures limited by necessarily low yields. Recently, the controlled folding of a long single DNA strand into an octahedron was reported¹⁴, an approach that may be thought of as 'single-stranded DNA origami'. The success of this work

suggested that the folding of long strands could, in principle, proceed without many misfoldings and avoid the problems of stoichiometry and purification associated with methods that use many short DNA strands.

I now present a versatile and simple 'one-pot' method for using numerous short single strands of DNA to direct the folding of a long, single strand of DNA into desired shapes that are roughly 100 nm in diameter and have a spatial resolution of about 6 nm. I demonstrate the generality of this method, which I term 'scaffolded DNA origami', by assembling six different shapes, such as squares, triangles and five-pointed stars. I show that the method not only provides access to structures that approximate the outline of any desired shape, but also enables the creation of structures with arbitrarily shaped holes or surface patterns composed of more than 200 individual pixels. The patterns on the 100-nm-sized DNA shapes thus have a complexity that is tenfold higher than that of any previously self-assembled arbitrary pattern and comparable to that achieved using AFM and STM surface manipulation⁴.

Design of scaffolded DNA origami

The design of a DNA origami is performed in five steps, the first two by hand and the last three aided by computer (details in Supplementary Note S1). The first step is to build a geometric model of a DNA structure that will approximate the desired shape. Figure 1a shows an example shape (outlined in red) that is 33 nm wide and 35 nm tall. The shape is filled from top to bottom by an even number of parallel double helices, idealized as cylinders. The helices are cut to fit the shape in sequential pairs and are constrained to be an integer number of turns in length. To hold the helices together, a periodic array of crossovers (indicated in Fig. 1a as small blue crosses) is incorporated; these crossovers designate positions at which strands running along one helix switch to an adjacent helix and continue there. The resulting model approximates the shape within one turn (3.6 nm) in the *x*-direction and roughly two helical widths (4 nm) in the

¹Departments of Computer Science and Computation & Neural Systems, California Institute of Technology, Pasadena, California 91125, USA.

y -direction. As noticed before in DNA lattices¹⁵, parallel helices in such structures are not close-packed, perhaps owing to electrostatic repulsion. Thus the exact y -resolution depends on the gap between helices. The gap, in turn, appears to depend on the spacing of crossovers. In Fig. 1a crossovers occur every 1.5 turns along alternating sides of a helix, but any odd number of half-turns may be used. In this study, data are consistent with an inter-helix gap of 1 nm for 1.5-turn spacing and 1.5 nm for 2.5-turn spacing, yielding a y -resolution of 6 or 7 nm, respectively.

Conceptually, the second step (illustrated in Fig. 1b) proceeds by folding a single long scaffold strand (900 nucleotides (nt) in Fig. 1b) back and forth in a raster fill pattern so that it comprises one of the two strands in every helix; progression of the scaffold from one helix to another creates an additional set of crossovers, the 'scaffold crossovers' (indicated by small red crosses in Fig. 1b). The fundamental constraint on a folding path is that the scaffold can form a crossover only at those locations where the DNA twist places it at a

tangent point between helices. Thus for the scaffold to raster progressively from one helix to another and onto a third, the distance between successive scaffold crossovers must be an odd number of half turns. Conversely, where the raster reverses direction vertically and returns to a previously visited helix, the distance between scaffold crossovers must be an even number of half-turns. Note that the folding path shown in Fig. 1b is compatible with a circular scaffold and leaves a 'seam' (a contour which the path does not cross).

Once the geometric model and a folding path are designed, they are represented as lists of DNA lengths and offsets in units of half-turns. These lists, along with the DNA sequence of the actual scaffold to be used, are input to a computer program. Rather than assuming 10.5 base pairs (bp) per turn (which corresponds to standard B-DNA twist), the program uses an integer number of bases between periodic crossovers (for example, 16 bp for 1.5 turns). It then performs the third step, the design of a set of 'staple strands' (the coloured DNA strands in Fig. 1c) that provide Watson–Crick complements for the

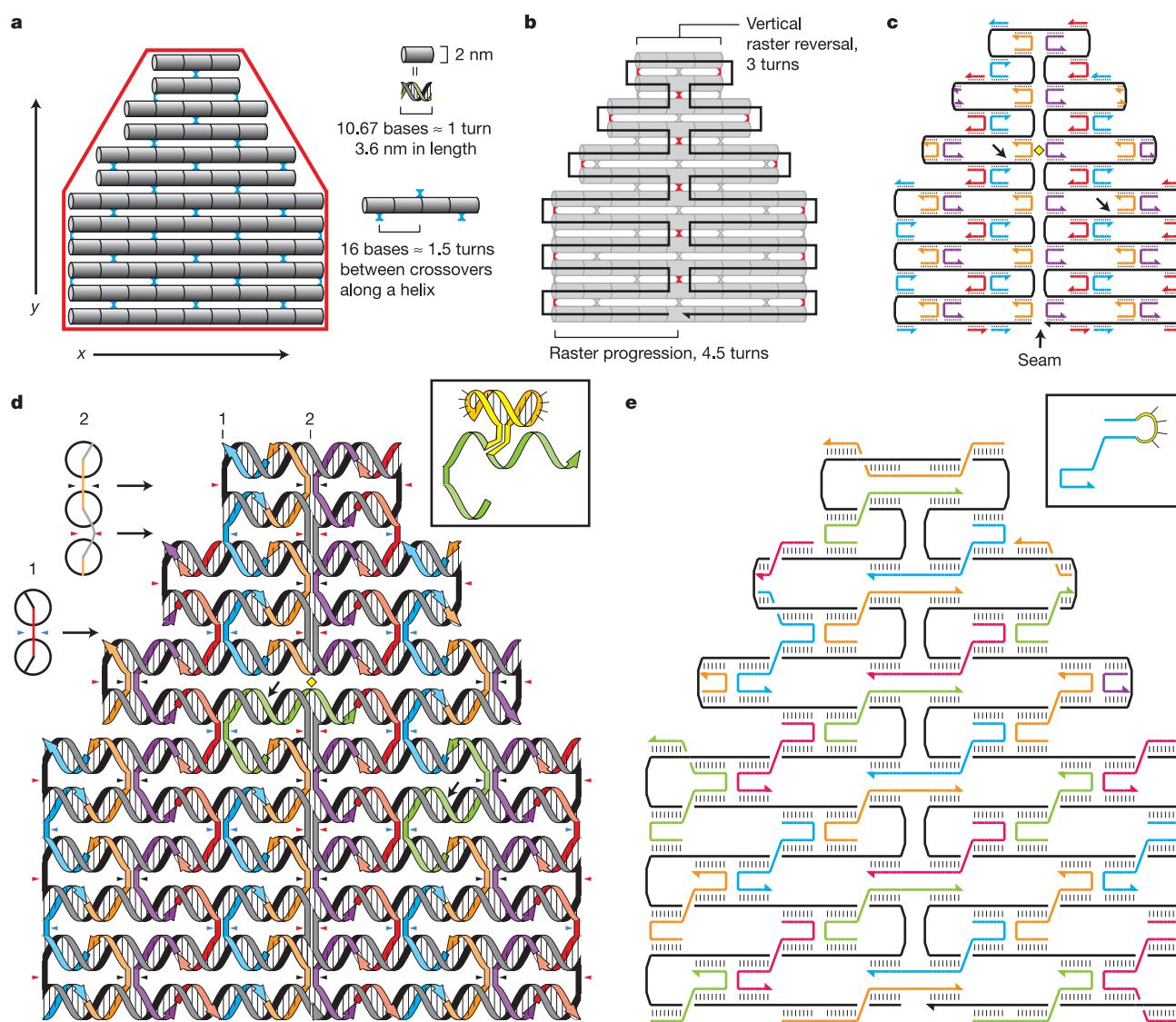


Figure 1 | Design of DNA origami. **a**, A shape (red) approximated by parallel double helices joined by periodic crossovers (blue). **b**, A scaffold (black) runs through every helix and forms more crossovers (red). **c**, As first designed, most staples bind two helices and are 16-mers. **d**, Similar to **c** with strands drawn as helices. Red triangles point to scaffold crossovers, black triangles to periodic crossovers with minor grooves on the top face of the shape, blue triangles to periodic crossovers with minor grooves on bottom. Cross-sections of crossovers (1, 2, viewed from left) indicate backbone positions

with coloured lines, and major/minor grooves by large/small angles between them. Arrows in **c** point to nicks sealed to create green strands in **d**. Yellow diamonds in **c** and **d** indicate a position at which staples may be cut and resealed to bridge the seam. **e**, A finished design after merges and rearrangements along the seam. Most staples are 32-mers spanning three helices. Insets show a dumbbell hairpin (**d**) and a 4-T loop (**e**), modifications used in Fig. 3.

scaffold and create the periodic crossovers. Staples reverse direction at these crossovers; thus crossovers are antiparallel, a stable configuration well characterized in DNA nanostructures¹⁶. Note that the crossovers in Fig. 1c are drawn somewhat misleadingly, in that single-stranded regions appear to span the inter-helix gap even though the design leaves no bases unpaired. In the assembled structures, helices are likely to bend gently to meet at crossovers so that only a single phosphate from each backbone occurs in the gap (as ref. 16 suggests for similar structures). Such small-angle bending is not expected to greatly affect the width of DNA origami (see also Supplementary Note S2).

The minimization and balancing of twist strain between crossovers is complicated by the non-integer number of base pairs per half-turn (5.25 in standard B-DNA) and the asymmetric nature of the helix (it has major and minor grooves). Therefore, to balance the strain¹⁵ caused by representing 1.5 turns with 16 bp, periodic crossovers are arranged with a glide symmetry, namely that the minor groove faces alternating directions in alternating columns of periodic crossovers (see Fig. 1d, especially cross-sections 1 and 2). Scaffold crossovers are not balanced in this way. Thus in the fourth step, the twist of scaffold crossovers is calculated and their position is changed (typically by a single bp) to minimize strain; staple sequences are recomputed accordingly. Along seams and some edges the minor groove angle (150°) places scaffold crossovers in tension with adjacent periodic crossovers (Fig. 1d, cross-section 2); such situations are left unchanged.

Wherever two staples meet there is a nick in the backbone. Nicks occur on the top and bottom faces of the helices, as depicted in Fig. 1d. In the final step, to give the staples larger binding domains with the scaffold (in order to achieve higher binding specificity and higher binding energy which results in higher melting temperatures), pairs of adjacent staples are merged across nicks to yield fewer, longer, staples (Fig. 1e). To strengthen a seam, an additional pattern of breaks and merges may be imposed to yield staples that cross the seam; a seam spanned by staples is termed 'bridged'. The pattern of merges is not unique; different choices yield different final patterns of nicks and staples. All merge patterns create the same shape but, as shown later, the merge pattern dictates the type of grid underlying any pixel pattern later applied to the shape.

Folding M13mp18 genomic DNA into shapes

To test the method, circular genomic DNA from the virus M13mp18 was chosen as the scaffold. Its naturally single-stranded 7,249-nt sequence was examined for secondary structure, and a hairpin with a 20-bp stem was found. Whether staples could bind at this hairpin was unknown, so a 73-nt region containing it was avoided. When a linear scaffold was required, M13mp18 was cut (in the 73-nt region) by digestion with *Bsr*BI restriction enzyme. While 7,176 nt remained available for folding, most designs did not fold all 7,176 nt; short (≤ 25 nt) 'remainder strands' were added to complement unused sequence. In general, a 100-fold excess of 200–250 staple and remainder strands were mixed with scaffold and annealed from

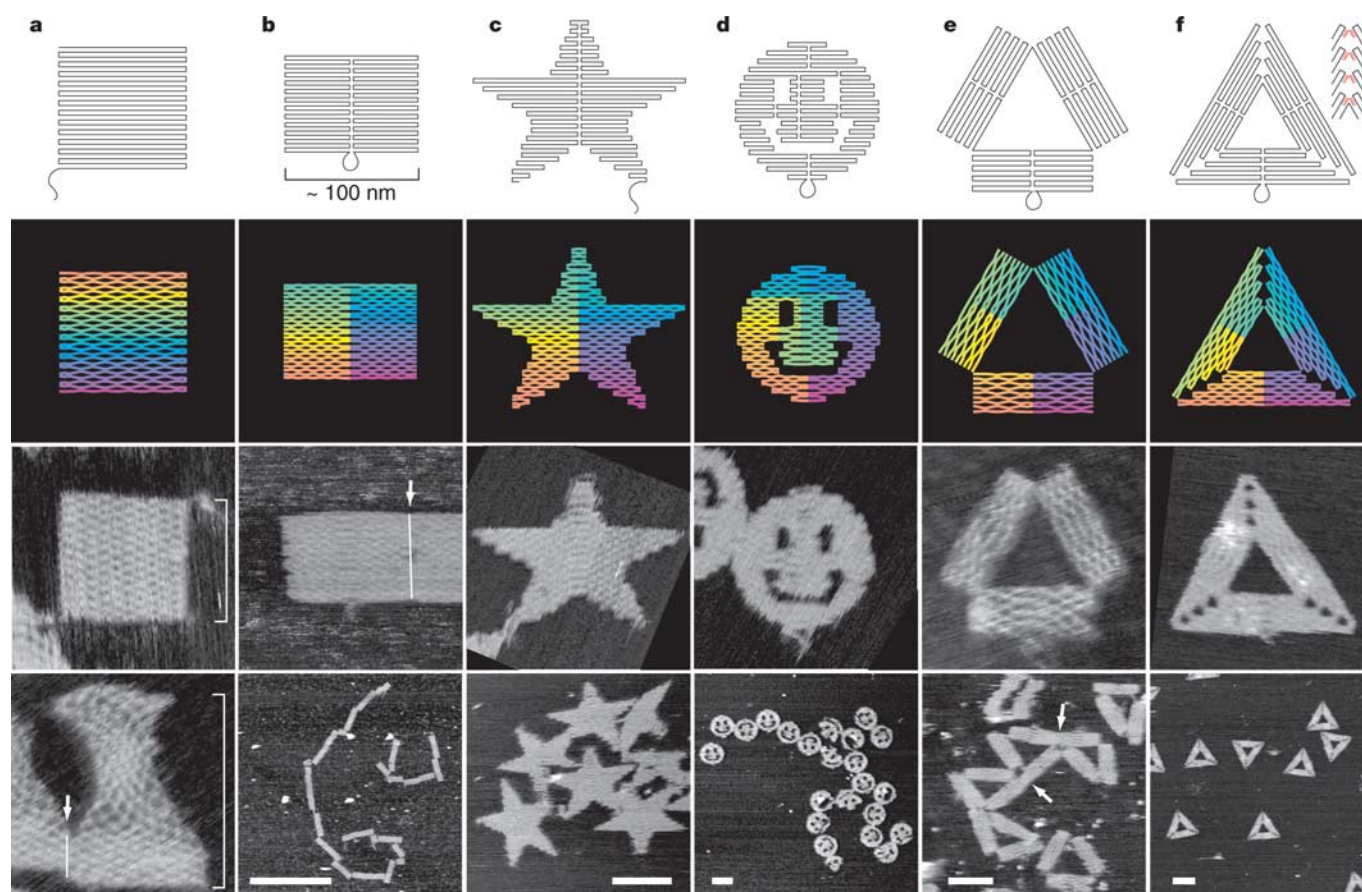


Figure 2 | DNA origami shapes. Top row, folding paths. **a**, square; **b**, rectangle; **c**, star; **d**, disk with three holes; **e**, triangle with rectangular domains; **f**, sharp triangle with trapezoidal domains and bridges between them (red lines in inset). Dangling curves and loops represent unfolded sequence. Second row from top, diagrams showing the bend of helices at crossovers (where helices touch) and away from crossovers (where helices bend apart). Colour indicates the base-pair index along the folding path; red

is the 1st base, purple the 7,000th. Bottom two rows, AFM images. White lines and arrows indicate blunt-end stacking. White brackets in **a** mark the height of an unstretched square and that of a square stretched vertically (by a factor >1.5) into an hourglass. White features in **f** are hairpins; the triangle is labelled as in Fig. 3k but lies face down. All images and panels without scale bars are the same size, 165 nm $\times$ 165 nm. Scale bars for lower AFM images: **b**, 1 μ m; **c-f**, 100 nm.

95 °C to 20 °C in <2 h. When samples were deposited on mica, only folded DNA structures stuck to the surface while excess staples remained in solution; AFM imaging thus proceeded under buffer without prior purification. Six different folds were explored; Fig. 2 gives their folding paths and their predicted and experimentally observed DNA structures. (Models and staple sequences are given in Supplementary Note S3, final designs appear in Supplementary Note S12. Experimental methods are given in Supplementary Note S4, results described here but not shown are in Supplementary Note S5.) Of the products imaged by AFM, a particular structure was considered qualitatively 'well-formed' if it had no defect (hole or indentation in the expected outline) greater than 15 nm in diameter. For each fold the fraction of well-formed structures, as a percentage of all distinguishable structures in one or more AFM fields, was calculated as a rough estimate of yield. I note that while some structures classified as well-formed had 15-nm defects, most had no defects greater than 10 nm in diameter.

First, a simple 26-helix square was designed (Fig. 2a). The square had no vertical reversals in raster direction, required a linear scaffold, and used 2.5-turn crossover spacing. Most staples were 26-mers that bound each of two adjacent helices as in Fig. 1c, but via 13 bases rather than 8. The design was made assuming a 1.5-nm inter-helix

gap; an aspect ratio of 1.05 (93.9 nm × 89.5 nm) was expected. By AFM, 13% of structures were well-formed squares (out of $S = 45$ observed structures) with aspect ratios from 1.00 to 1.07 and bore the expected pattern of crossovers (Fig. 2a, upper AFM image). Of the remaining structures, ~25% were rectangular fragments, and ~25% had an hourglass shape that showed a continuous deformation of the crossover lattice (Fig. 2a, lower AFM image). Sequential imaging documented the stretching of a square into an hourglass, suggesting that hourglasses were originally squares that stretched upon deposition or interaction with the AFM tip. No subsequent designs exhibited stretching. Other designs had either a tighter 1.5-turn spacing with 32-mer staples spanning three helical domains (Fig. 2b–d, f) or smaller domains that appeared to slide rather than stretch (Fig. 2e).

To test the formation of a bridged seam, a rectangle was designed (Fig. 2b) according to the scheme outlined in Fig. 1e using 1.5-turn crossover spacing, 32-mer staples and a circular scaffold. As seen in Fig. 2b, the central seam and associated pattern of crossovers was easily visualized (upper AFM image). Rectangles stacked along their vertical edges, often forming chains up to 5 μm long (lower AFM image). The yield of well-formed rectangles was high (90%, $S = 40$), and so rectangles were used to answer basic questions concerning inter-helix gaps, base-stacking, defects and stoichiometry. AFM drift

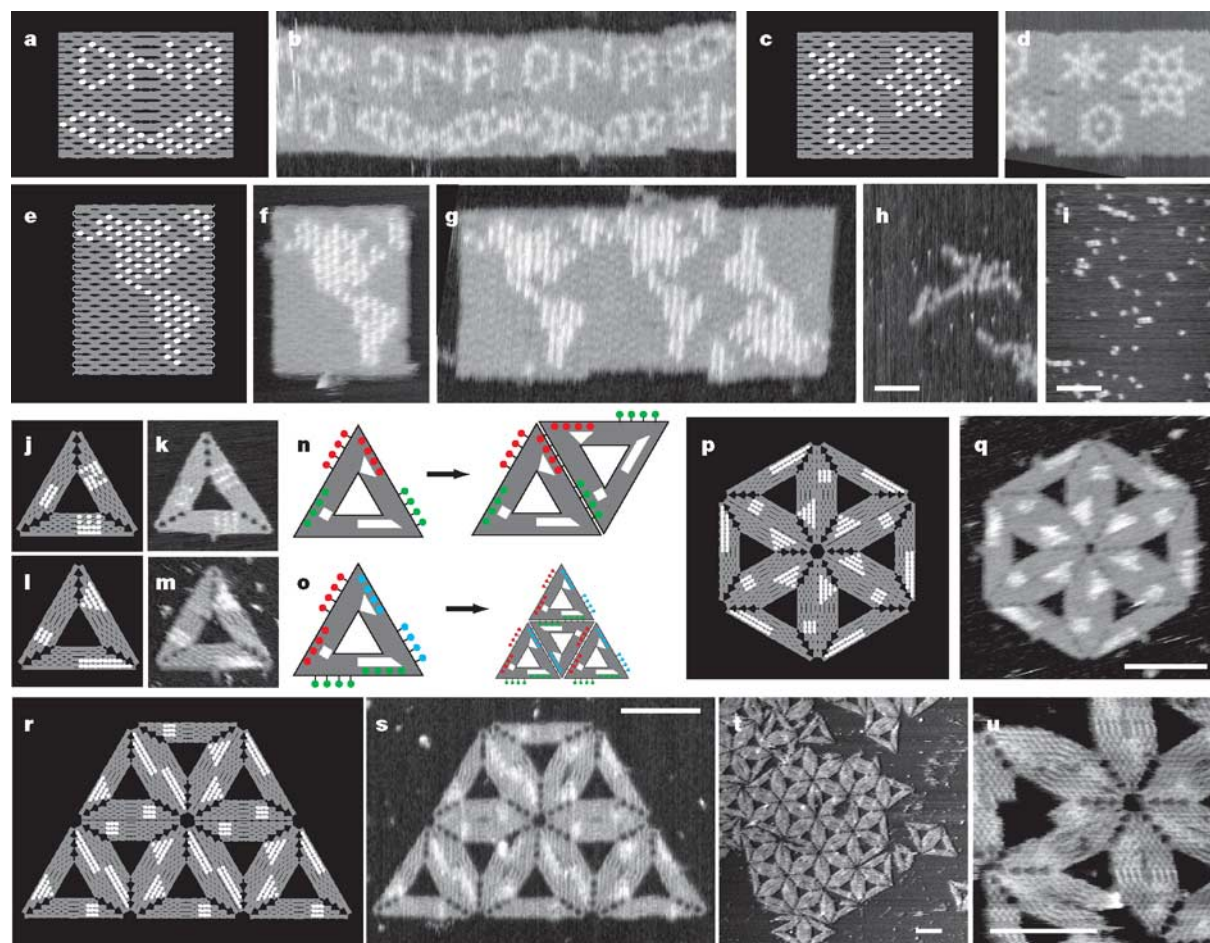


Figure 3 | Patterning and combining DNA origami. **a**, Model for a pattern representing DNA, rendered using hairpins on a rectangle (Fig. 2b). **b**, AFM image. One pixelated DNA turn (~100 nm) is 30× the size of an actual DNA turn (~3.6 nm) and the helix appears continuous when rectangles stack appropriately. Letters are 30 nm high, only 6× larger than those written using STM in ref. 3; 50 billion copies rather than 1 were formed. **c**, **d**, Model and AFM image, respectively, for a hexagonal pattern that highlights the nearly hexagonal pixel lattice used in **a**–**i**. **e**–**i**, Map of the western hemisphere, scale 1:2 × 10¹⁴, on a rectangle of different aspect ratio. Normally such rectangles aggregate (**h**) but 4-T loops or tails on edges (white

lines in **e**) greatly decrease stacking (**i**). **j**–**m**, Two labellings of the sharp triangle show that each edge may be distinguished. In **j**–**u**, pixels fall on a rectilinear lattice. **n**–**u**, Combination of sharp triangles into hexagons (**n**, **p**, **q**) or lattices (**o**, **r**–**u**). Diagrams (**n**, **o**) show positions at which staples are extended (coloured protrusions) to match complementary single-stranded regions of the scaffold (coloured holes). Models (**p**, **r**) permit comparison with data (**q**, **s**). The largest lattice observed comprises only 30 triangles (**t**). **u** shows close association of triangles (and some breakage). **d** and **f** were stretched and sheared to correct for AFM drift. Scale bars: **h**, **i**, 1 μm; **q**, **s**–**u**, 100 nm.

often distorts aspect ratios so that inter-helix gaps cannot be inferred from the aspect ratio of a single rectangle. A range of aspect ratios implied a gap size from 0.9 to 1.2 nm; later designs assume 1 nm. Whatever the exact value, it is consistent: aspect ratios were invariant along stacked chains with dozens of rectangles. Such stacking was almost completely abolished by omitting staples along vertical edges. On the other hand, stacking across the seam of an unbridged rectangle (as in Fig. 1c) kept 65% of structures ($S = 40$) well-formed; the rest showed some degree of dislocation at the seam. Other defects, such as the intentional omission of single staples, could be visualized as 5–10-nm holes. However, sharp tips and high tapping amplitudes were required; repeated scanning created holes difficult to distinguish from holes due to missing strands. This effect also increased uncertainty when stoichiometry was varied. When staple excesses of approximately 100:1 and 9:1 were used, the frequencies of 5–10-nm holes (a few per rectangle) were indistinguishable. At 2:1, rectangles were similar; perhaps a greater fraction were malformed. At 1.5:1, rectangles formed but had holes up to $\sim 10\%$ of their area in size. At a 1:1 ratio, $<1\%$ of structures were rectangular.

To demonstrate the creation of arbitrary shapes, a five-pointed star was designed with 1.5-turn spacing, 32-mer staples and a linear rather than circular scaffold (Fig. 2c). Designed assuming a 1.5-nm inter-helix gap (the work was carried out before the gap for 1.5-turn spacing was measured), the stars are somewhat squat (Fig. 2c, upper AFM image). Still, the stars show that the width of a shape may be approximated to within one DNA turn. Many of the structures observed were star fragments (Fig. 2c, lower AFM image), and only 11% ($S = 70$) were well-formed. The low yield of stars (and squares, see above) may be due to strand breakage occurring during *Bsr*BI digestion or subsequent steps to remove the enzyme; when untreated circular scaffold was folded into stars, 63% ($S = 43$) were well-formed. To show that DNA origami need not be topological disks, and that scaffolds can be routed arbitrarily through shapes, a three-hole disk was designed (Fig. 2d). Although the shape approximated is symmetric, the folding path is highly asymmetric and has five distinct seams. Unlike the rectangles, which rarely break or fold, three-hole disks exhibit several characteristic deformations (Fig. 2d, lower AFM image); still, 70% ($S = 90$) were well-formed.

DNA origami is not limited to the approximation of shapes by raster fill: some shapes can be created more exactly by combining distinct raster fill domains in non-parallel arrangements. Figure 2e shows a triangle built from three separate, 2.5-turn spacing rectangular domains; only single covalent bonds along the scaffold hold the domains together. But the desired equiangular triangles (upper AFM image) were rarely observed ($<1\%$, $S = 199$). As seen in the lower AFM image, stacking caused rectangular domains of separate triangles to bind; this effect and the flexibility of the single-bond joints at the vertices may account for the ease with which these triangles deform. To solve these problems, 'sharp triangles', built from trapezoidal domains with 1.5-turn spacing, were designed (Fig. 2f). The slanted edges of the trapezoids meet at the triangle vertices and allow the addition of bridging staples along these interfaces. Sharp triangles remained separated and equiangular (Fig. 2f, lower AFM image); 88% were well-formed ($S = 78$). Even when bridging staples at the vertices were not used, a large number of sharp triangles were well-formed (55%, $S = 22$). These 'weakened' sharp triangles provided the most stringent test of the estimated inter-helix gap, because too high or low an estimate would have caused gaps or overlaps between trapezoids. Gaps of 10 nm occasionally appeared but overlaps were never observed, suggesting that 1 nm may be a slight underestimate of the inter-helix gap.

Patterning and combining DNA origami

In addition to binding the DNA scaffold and holding it in shape, staple strands provide a means for decorating shapes with arbitrary patterns of binary pixels. Given a shape, the original set of staples is taken to represent binary '0's; a new set of labelled staples, one for

each original staple, is used to represent binary '1's. Patterns are created by mixing appropriate subsets of these strands. In this way, any desired pattern can be made.

In principle, a variety of DNA modifications—for example, biotin or fluorophores—could serve as labels. Here, 'dumbbell hairpins' (Fig. 1d inset, Supplementary Note S6), designed to avoid dimerization at high concentration, were added to the middle of 32-mer staples at the position of merges made during design. Depending on the merge pattern, the resulting pixel pattern was either rectilinear, with adjacent columns of hairpins on alternate faces of the shape, or staggered and nearly hexagonally packed, with all hairpins on the same face. In AFM images labelled staples give greater height contrast (3 nm above the mica) than unlabelled staples (~ 1.5 nm), which results in a pattern of light '1' and dark '0' pixels. Several patterns (Fig. 3), each with ~ 200 pixels, illustrate the generality of this technique.

Yields of patterned origami were similar to those of unpatterned origami; for the pattern in Fig. 3a, 91% ($S = 85$) of rectangles were well-formed. Because rectilinear patterns imaged poorly, only staggered patterns were examined quantitatively. Distances measured between pairs of '1' pixels in alternating columns (two pixel widths: 11.5 ± 0.9 nm, mean $\pm$ s.d., $n = 26$) and adjacent rows (one pixel height: 6.6 ± 0.5 nm, $n = 24$) are consistent with the theoretically expected pixel size of 5.4 nm $\times$ 6 nm. Most defects take the form of 'missing pixels'; that is, pixels that should image as '1's but image as '0's instead. 94% of '1' pixels (of 1,080 observed) were visualized. Whether missing pixels represent real defects or artefacts of imaging is unknown; sequential AFM images occasionally showed '1' pixels that later converted irreversibly to '0' pixels, suggesting tip-induced damage. Stoichiometric errors, synthetic errors, or unwanted secondary structure are not implicated for any particular strand, as the position of missing pixels appeared random (Fig. 3b, f and g).

Stacking of shapes along blunt-ended helices provides an uncontrolled mechanism for the creation of larger structures (Fig. 3b). Instead of removing staples on the edge of a rectangle to avoid stacking (as described previously), 4-T hairpin loops (four thymines in a row, Fig. 1e, inset) or 4-T tails can be added to edge staples (Fig. 3e, f); stacked chains of 3–5 rectangles still formed (Fig. 3g), but 30% of rectangles ($S = 319$) occurred as monomers (Fig. 3i). Without hairpins, all rectangles occurred in aggregates (Fig. 3h).

Controlled combination of shapes was achieved by designing 'extended staples' that connected shapes along their edges. To create a binding interaction between two particular edges, extended staples were designed by merging and breaking normal staples along these edges (Supplementary Note S7). Starting with sharp triangles, this approach was used to create finite hexagons (Fig. 3n, p, q) as well as periodic structures (triangular lattice; Fig. 3o, r–u). I note that the successful combination of shapes (unlike the successful formation of individual shapes) is in principle very sensitive to the concentrations of extended staples (which should ideally be equal to that of the scaffold). Poor stoichiometry may play a role in the poor yield of hexagons ($<2\%$, $S = 70$) and lattices (not measured).

Discussion

The scaffolded self-assembly of DNA strands has been used to create linear structures^{17,18} and proposed as a method for creating arbitrary patterns^{18,19}. But the widespread use of scaffolded self-assembly, and in particular the use of long DNA scaffolds in combination with hundreds of short strands, has been inhibited by several misconceptions: it was assumed that (1) sequences must be optimized²⁰ to avoid secondary structure or undesired binding interactions, (2) strands must be highly purified, and (3) strand concentrations must be precisely equimolar. These three criteria are important for the formation of many DNA nanostructures and yet all three are ignored in the present method. For example, M13mp18 is essentially a natural sequence that has a predicted secondary structure which is more stable (lower in energy) than similar random sequences

(Supplementary Note S8). Further, stocks of staples each contained a few per cent truncation products, stock concentrations were measured with at least 10% error, and staples were used successfully at stoichiometries that varied over an order of magnitude.

I suggest that several factors contribute to the success of scaffolded DNA origami (even though the method ignores the normal, careful practices of DNA nanotechnology). These are (1) strand invasion, (2) an excess of staples, (3) cooperative effects and (4) design that intentionally does not rely on binding between staples. Briefly (details are given in Supplementary Note S9), strand invasion may allow correct binding of excess full-length staples to displace unwanted secondary structure, incorrect staples, or grossly truncated staples. Further, each correct addition of a staple organizes the scaffold for subsequent binding of adjacent staples and precludes a large set of undesired secondary structures. Last, because staples are not designed to bind one another, their relative concentrations do not matter.

The method presented here is easy to implement, high yield and relatively inexpensive. Three months of effort went into the design program. In addition, each structure required about one week to design and one week to synthesize (commercially); the mixing and annealing of strands required a few hours. The greatest experimental difficulty was acquiring high-resolution AFM images, typically taking two days per structure. For rigid designs using circular scaffolds (rectangles with patterns, three-hole disks, and sharp triangles), yields of qualitatively well-formed structures were at least 70%. A better understanding of folding will depend on less-destructive imaging and quantification of small (<15 nm) defects. A possible objection to the routine use of the method is the potential cost of staples; unlike the scaffold, staples cannot be cloned. However, unpurified strands are inexpensive so that the scaffold constitutes 80% of the cost, even when using a 100-fold excess of staples (Supplementary Note S10).

I believe that scaffolded DNA origami can be adapted to create more complex or larger structures. For example, the design of three-dimensional structures should be accessible using a straightforward adaptation of the raster fill method given here. If non-repetitive scaffolds of megabase length can be prepared, micrometre-size origami with 20,000 features may be possible. However, the requirement for unique sequence information means that the method cannot be scaled up arbitrarily; whenever structures above a critical size or level of complexity are desired, it will therefore be necessary to combine scaffolded DNA origami with hierarchical self-assembly^{10,11}, algorithmic self-assembly²², or top-down fabrication techniques.

An obvious application of patterned DNA origami would be the creation of a 'nanobreadboard', to which diverse components could be added. The attachment of proteins²³, for example, might allow novel biological experiments aimed at modelling complex protein assemblies and examining the effects of spatial organization, whereas molecular electronic or plasmonic circuits might be created by attaching nanowires, carbon nanotubes or gold nanoparticles²⁴. These ideas suggest that scaffolded DNA origami could find use in fields as diverse as molecular biology and device physics.

Received 7 September 2005; accepted 12 January 2006.

1. Feynman, R. P. There's plenty of room at the bottom. *Engineering and Science* 23(5), 22–36 (Caltech, February, 1960).

2. Junno, T., Deppert, K., Montelius, L. & Samuelson, L. Controlled manipulation of nanoparticles with an atomic force microscope. *Appl. Phys. Lett.* 66, 3627–3629 (1995).
3. Eigler, D. M. & Schweizer, E. K. Positioning single atoms with a scanning tunnelling microscope. *Nature* 344, 524–526 (1990).
4. Heinrich, A. J., Lutz, C. P., Gupta, J. A. & Eigler, D. M. Molecular cascades. *Science* 298, 1381–1387 (2002).
5. Whitesides, G. M., Mathias, J. P. & Seto, C. T. Molecular self-assembly and nanochemistry: a chemical strategy for the synthesis of nanostructures. *Science* 254, 1312–1319 (1991).
6. Yokoyama, T., Yokoyama, S., Kamikado, T., Okuno, Y. & Mashiko, S. Self-assembly on a surface of supramolecular aggregates with controlled size and shape. *Nature* 413, 619–621 (2001).
7. Mao, C. B. et al. Virus-based toolkit for the directed synthesis of magnetic and semiconducting nanowires. *Science* 303, 213–217 (2004).
8. Seeman, N. C. Nucleic-acid junctions and lattices. *J. Theor. Biol.* 99, 237–247 (1982).
9. Seeman, N. C. & Lukeman, P. S. Nucleic acid nanostructures: bottom-up control of geometry on the nanoscale. *Rep. Prog. Phys.* 68, 237–270 (2005).
10. Chworos, A. et al. Building programmable jigsaw puzzles with RNA. *Science* 306, 2068–2072 (2004).
11. Park, S. H. et al. Finite-size, fully-addressable DNA tile lattices formed by hierarchical assembly procedures. *Angew. Chem.* 118, 749–753 (2006).
12. Chen, J. & Seeman, N. C. The synthesis from DNA of a molecule with the connectivity of a cube. *Nature* 350, 631–633 (1991).
13. Zhang, Y. & Seeman, N. C. The construction of a DNA truncated octahedron. *J. Am. Chem. Soc.* 116, 1661–1669 (1994).
14. Shih, W. M., Quispe, J. D. & Joyce, G. F. A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature* 427, 618–621 (2004).
15. Rothmund, P. W. K. et al. Design and characterization of programmable DNA nanotubes. *J. Am. Chem. Soc.* 126, 16344–16353 (2004).
16. Fu, T.-J. & Seeman, N. C. DNA double-crossover molecules. *Biochemistry* 32, 3211–3220 (1993).
17. LaBean, T. H., Winfree, E. & Reif, J. H. in *DNA Based Computers V* (eds Winfree, E. & Gifford, D. K.) 123–140 (Vol. 54 of DIMACS, AMS Press, Providence, Rhode Island, 1999).
18. Yan, H., LaBean, T. H., Feng, L. & Reif, J. H. Directed nucleation assembly of DNA tile complexes for barcode-patterned lattices. *Proc. Natl Acad. Sci. USA* 100, 8103–8108 (2003).
19. Reif, J. H. in *Proc. 29th Int. Colloquium on Automata, Languages, and Programming (ICALP)* (eds Widmayer, P., Ruiz, F. T., Bueno, R. M., Hennessy, M., Eidenbenz, S. & Conejo, R.) 1–21 (Vol. 2380 of Lecture Notes in Computer Science, Springer, New York, 2002).
20. Seeman, N. C. De novo design of sequences for nucleic acid structural engineering. *J. Biomol. Struct. Dyn.* 8, 573–581 (1990).
21. Winfree, E. in *DNA Based Computers* (eds Lipton, R. J. & Baum, E. B.) 199–221 (Vol. 27 of DIMACS, AMS Press, Providence, Rhode Island, 1996).
22. Rothmund, P. W. K., Papadakis, N. & Winfree, E. Algorithmic self-assembly of DNA Sierpinski triangles. *PLoS Biol.* 2, e424 (2004).
23. Yan, H., Park, S. H., Finkelstein, G., Reif, J. H. & LaBean, T. H. DNA-templated self-assembly of protein arrays and highly conductive nanowires. *Science* 301, 1882–1884 (2003).
24. Le, J. D. et al. DNA-templated self-assembly of metallic nanocomponent arrays on a surface. *Nano Lett.* 4, 2343–2347 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements I thank E. Winfree for discussions and providing a stimulating laboratory environment; B. Yurke for the term 'nanobreadboard'; N. Papadakis, L. Adleman, J. Goto, R. Barish, R. Schulman, R. Hariadi, M. Cook and M. Diehl for discussions; B. Shaw for a gift of AFM tips; A. Schmidt for coordinating DNA synthesis; and K. Yong, J. Crouch and L. Hein for administrative support. This work was supported by National Science Foundation Career and Nano grants to E. Winfree as well as fellowships from the Beckman Foundation and Caltech Center for the Physics of Information.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to P.W.K.R. (pwkr@dna.caltech.edu).

Robust *Salmonella* metabolism limits possibilities for new antimicrobials

Daniel Becker^{1*}, Matthias Selbach^{2*}, Claudia Rollenhagen^{1*}, Matthias Ballmaier³, Thomas F. Meyer¹, Matthias Mann² & Dirk Bumann^{1,4}

New antibiotics are urgently needed to control infectious diseases. Metabolic enzymes could represent attractive targets for such antibiotics, but *in vivo* target validation is largely lacking. Here we have obtained *in vivo* information about over 700 *Salmonella enterica* enzymes from network analysis of mutant phenotypes, genome comparisons and *Salmonella* proteomes from infected mice. Over 400 of these enzymes are non-essential for *Salmonella* virulence, reflecting extensive metabolic redundancies and access to surprisingly diverse host nutrients. The essential enzymes identified were almost exclusively associated with a small subgroup of pathways, enabling us to perform a nearly exhaustive screen. Sixty-four enzymes identified as essential in *Salmonella* are conserved in other important human pathogens, but almost all belong to metabolic pathways that are inhibited by current antibiotics or that have previously been considered for antimicrobial development. Our comprehensive *in vivo* analysis thus suggests a shortage of new metabolic targets for broad-spectrum antibiotics, and draws attention to some previously known but unexploited targets.

Microbial resistance to antibiotics has become a serious threat for human health¹. To ensure effective management of infectious diseases, new antibiotics that inhibit different bacterial targets are urgently needed. Metabolic enzymes are particularly attractive drug target candidates owing to their central role in microbial physiology, their high conservation among various pathogens, and the suitability of enzyme assays for high-throughput identification of hit compounds and lead optimization². Currently used antibiotics target only a small number of metabolic enzymes, suggesting that numerous additional targets might exist.

Indeed, experimental *in vitro* studies³ as well as theoretical genome-scale network topology⁴ and computational models of metabolic fluxes⁵ have revealed hundreds of important bacterial enzymes. However, which enzymes are identified as essential depends on the growth conditions^{6,7}. As laboratory cultures do not perfectly mimic the conditions in infected hosts, *in vivo* experiments are needed to identify metabolic drug targets⁸; however, systematic studies of pathogen metabolism in infected hosts are lacking. *Salmonella enterica* is a particularly suitable model pathogen for *in vivo* studies as it causes major human diseases (enteritis and systemic typhoid fever^{9,10}) that are mimicked in well-established small animal models of infection^{11,12}. Moreover, the extensively characterized metabolic networks of *Salmonella* and closely related *Escherichia coli* provide a good framework for data interpretation.

A model of *Salmonella* metabolism *in vivo*

To reconstruct *Salmonella* metabolism during systemic infection, we first analysed published phenotypes of *Salmonella* metabolic mutants in a mouse model of typhoid fever (Supplementary Table 1). Enzymes were classified as (1) 'essential' if their inactivation resulted in more than 1,000-fold attenuation compared to wild-type *Salmonella* (for details see Supplementary Methods), (2) 'contributing to virulence' if their inactivation partially decreased virulence, or (3) 'dispensable' if

their inactivation had no significant effect on virulence. We also analysed clinical data on three experimental vaccine strains of *Salmonella enterica* serovar Typhi carrying defined metabolic mutations (Supplementary Table 1). These strains are attenuated but still cause fever and/or bacteraemia in human volunteers, indicating that the mutated enzymes contribute to virulence but are not essential. In contrast, we took the clinical efficacy of trimethoprim, sulphonamides and cephalosporin antibiotics against human typhoid fever as evidence of essential *in vivo* roles for the respective *Salmonella* enzymes they inhibit (Supplementary Table 1).

To obtain additional information regarding dispensable enzymes, we compared the genome sequences of four *Salmonella enterica* serovars (Typhimurium¹³, Enteritidis¹⁴, Typhi^{15,16} and Paratyphi A¹⁷) that cause typhoid fever-like diseases in various hosts. This analysis revealed 98 metabolic genes that are present in serovar Typhimurium (our reference serovar) but are non-functional in the other serovars (pseudogenes or missing genes; see Supplementary Table 2), suggesting that these genes are dispensable for systemic virulence.

Our combined data revealed functional information about 196 enzymes (15 essential, 43 contributing and 138 dispensable; Supplementary Table 3). To obtain indirect information regarding additional enzymes with related biochemical functions, we generated a graphical metabolic network model for *Salmonella*, adapted from the EcoCyc model for *E. coli* (available at <http://www.ecocyc.org>)¹⁸, using the *Salmonella enterica* serovar Typhimurium LT2 genome sequence¹³ and information available in the manually curated and continuously updated EcoCyc and EcoSal (<http://www.ecosal.org>) databases (Fig. 1; see Supplementary Methods for model generation). *Salmonella* enzymes were labelled according to their *in vivo* relevance (essential enzymes in red, enzymes contributing to virulence in orange, and dispensable enzymes in blue). Indirect evidence for the function of additional enzymes was obtained by assuming that all

¹Max-Planck-Institute for Infection Biology, Department of Molecular Biology, D-10117 Berlin, Germany. ²Max-Planck-Institute of Biochemistry, Department of Proteomics and Signal Transduction, D-82152 Martinsried, Germany. ³Hannover Medical School, Flow Cytometry Core Facility, D-30625 Hannover, Germany. ⁴Hannover Medical School, Institute of Immunology, 'Mucosal Infections' Junior Research Group OE 9421, D-30625 Hannover, Germany.

*These authors contributed equally to this work.

non-redundant enzymes involved in pathways containing essential enzymes were also essential ('indirect evidence' column in Supplementary Table 3; white labels in Fig. 1). We similarly derived indirect evidence for enzymes contributing to virulence and dispensable enzymes (see Supplementary Methods for details). Overall, this network analysis revealed direct and indirect functional information about 455 enzymes in the typhoid fever model (95 essential, 91 contributing and 269 dispensable; Supplementary Table 3). For the enteritis model¹¹, a similar analysis of the little available data revealed 31 dispensable but no functionally relevant enzymes (Supplementary Table 3 and Supplementary Fig. 1).

More than 250 genes have previously been reported to be essential for *in vitro* growth of *Salmonella* in Luria-Bertani medium¹⁹, and many of these might also be essential during infection. However, independent published evidence has suggested that at least 46 of these reportedly essential genes might actually be non-essential, as gene inactivation resulted in viable mutants that grew in Luria-Bertani medium (Supplementary Table 4). We therefore did not classify these genes as essential in our metabolic reconstruction.

In vivo proteomics in *Salmonella*

To obtain complementary data on *Salmonella* metabolism during infection *in vivo*, we investigated *Salmonella* enzyme expression in infected mice. Such data has not previously been available for any pathogen, as detecting very low levels of pathogen proteins in the

presence of a large excess of host protein is technically difficult. We achieved this task by purifying live *Salmonella* expressing green fluorescent protein (GFP) from infected mouse tissues using flow cytometry (Fig. 2; see Methods). *Salmonella* enzymes were then identified by mass spectrometry-based proteomics²⁰.

Salmonella sorted from spleen homogenates (mouse typhoid fever model) were contaminated with host proteins (which made up about 70% of the identified proteins; data not shown) but we were able to detect 12 out of 21 *Salmonella* transfer RNA synthetases¹³, and 45 out of 53 ribosomal proteins (Supplementary Table 5), indicating that about half of the abundant *Salmonella* proteins were detected. Most low-abundance *Salmonella* proteins probably escaped detection because of the limited amount and high complexity of the sample. Material sorted from the caecum (enteritis model¹¹) contained about 75% *Salmonella* proteins, with *Bacteroides* proteins as the dominant source of contamination (data not shown). The identification of 20 *Salmonella* tRNA synthetases and 51 ribosomal subunits (Supplementary Table 5) indicated a high detection rate of abundant *Salmonella* proteins.

In vivo *Salmonella* proteomes from both mouse models contained high numbers of metabolic enzymes (228 enzymes among 370 identified *Salmonella* proteins in the typhoid fever model; 539 enzymes among 835 identified *Salmonella* proteins in the enteritis model; Supplementary Tables 3 and 5). Detected enzymes are shown in yellow in the respective metabolic models, revealing

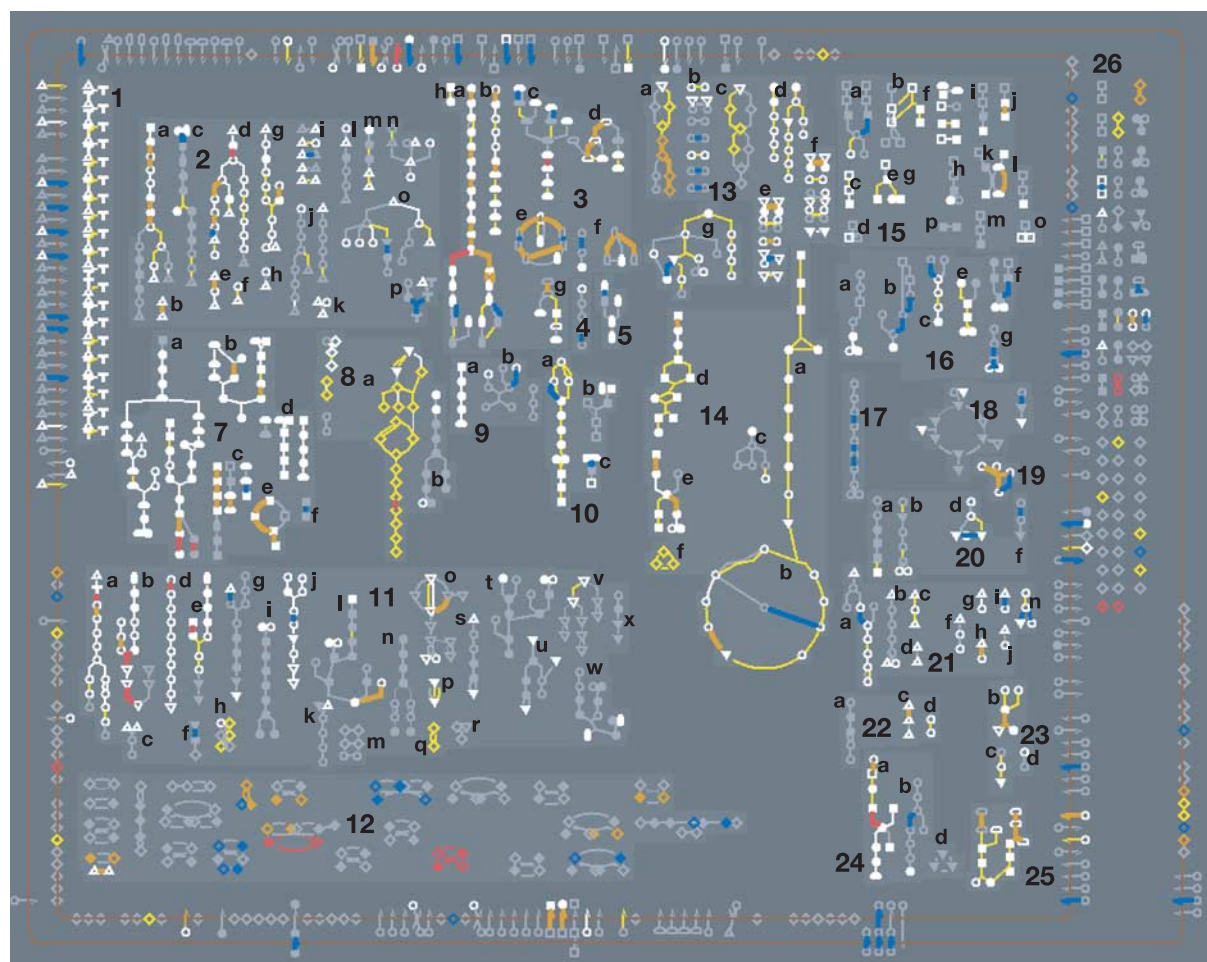


Figure 1 | Overview of *Salmonella* metabolism during typhoid fever. Symbols represent metabolites and connecting lines represent enzymes catalysing the corresponding reactions. Enzyme expression detected *in vivo* (yellow) and *in vivo* enzyme relevance (essential enzymes in red, enzymes contributing to virulence in orange, and dispensable enzymes

in blue), indirectly inferred enzymes and metabolites (white), and metabolites and enzymes with no available *in vivo* evidence (grey) are shown. Brown lines represent the two *Salmonella* membranes. Pathway numbers correspond to Supplementary Table 3.

comprehensive coverage of many functionally relevant pathways (Fig. 1 and Supplementary Fig. 1).

Technical limitations of flow cytometry and mass spectrometry techniques resulted in incomplete coverage of the *Salmonella* enzyme network (see above). To provisionally fill in some of these gaps, we used indirect evidence derived from the metabolic network model. In particular, detection of an enzyme involved in only one specific pathway was taken as a hint for the expression of other non-redundant enzymes within that same pathway (Supplementary Table 3; white labels in Fig. 1 and Supplementary Fig. 1). Direct and indirect evidence suggests the expression of 410 *Salmonella* metabolic enzymes in the typhoid fever model, including 233 enzymes for which no other *in vivo* information is available (Supplementary Table 3). A combination of functional and proteome data revealed *in vivo* information for a total of 688 *Salmonella* enzymes in the typhoid fever model (Fig. 3a and Supplementary Tables 3 and 5). A similar analysis for the enteritis model revealed *in vivo* information for 744 *Salmonella* enzymes (Supplementary Tables 2, 3 and 5).

Common properties of essential enzymes

Our overview of *Salmonella* metabolic pathways (Fig. 1 and Supplementary Fig. 1) reveals a complex picture of essential, contributing and dispensable enzymes, with no obvious simple patterns. To determine whether there are any underlying common properties of essential or dispensable enzymes, we considered the typhoid fever model in more detail, characterizing 196 of the enzymes with experimentally determined phenotypes.

To have a role in virulence, an enzyme must be expressed during infection. Indeed, we found that most known essential enzymes and more than half of the contributing enzymes belonged to pathways detected in the *Salmonella in vivo* proteome, whereas only a small minority of dispensable enzymes was detected (Fig. 3b and Supplementary Table 3). These data indicate a strong association between detectable enzyme expression and functional relevance compared to other well-characterized systems²¹. However, some 40% of enzymes that are known to have a role in typhoid fever

models escaped detection because of limited material and high sample complexity (see above).

Essential enzymes could have distinct metabolic functions. To test this hypothesis, we classified the 196 enzymes into six different pathway categories (Fig. 3c and Supplementary Tables 1 and 2). As we expected, all enzymes with known redundancy (isozymes or alternative pathways) were non-essential in the typhoid fever model. All characterized transporters and catabolic enzymes except for NagA were also non-essential. NagA deficiency impairs *N*-acetylglucosamine utilization, but avirulence as a result of NagA deficiency is more likely due to indirect inhibition of lipopolysaccharide biosynthesis²². In contrast to other functional categories, several biosynthetic *Salmonella* enzymes were essential for systemic virulence (Fig. 3c). This was particularly true for the biosynthesis of products that *Salmonella* cannot efficiently obtain from its host ('group A' enzymes) because (1) host tissues do not contain these metabolites, (2) *Salmonella* lack appropriate high-affinity uptake systems, and/or (3) *Salmonella* inefficiently funnel transported metabolites into the relevant pathways. In contrast to group A biosynthetic enzymes, no enzyme involved in the biosynthesis of readily used metabolites ('group B') was essential. Finally, two non-redundant enzymes involved in central intermediary metabolism and energy production were also essential. Collectively, these data suggest that essential *Salmonella* enzymes are almost exclusively found in a small, distinct subgroup of pathways.

New essential *Salmonella* enzymes

The *Salmonella* genome encodes more than 2,200 proteins with putative metabolic functions¹³, most of which remain uncharacterized. Complete *in vivo* evaluation would be prohibitively cost- and labour-intensive, but we could narrow down the number of candidate target enzymes to a more manageable number based on our observations that a small subgroup of non-redundant biochemical pathways probably contains most of the essential metabolic enzymes (Fig. 3c).

A genome-wide search for non-redundant enzymes involved in group A biosynthesis, central intermediary metabolism or energy

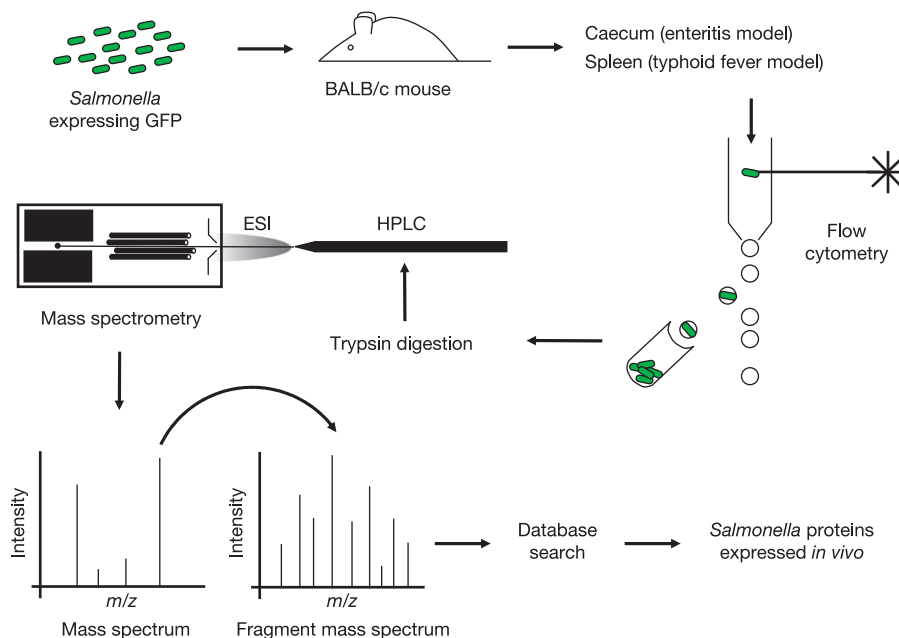


Figure 2 | Experimental strategy for *in vivo* *Salmonella* proteomics. Mice were infected with *Salmonella* expressing GFP. After several days, fluorescent GFP-expressing *Salmonella* were purified from spleen or caecum homogenates by flow cytometry. Purified *Salmonella* were digested with trypsin and the resulting peptide mixtures were separated by high-

performance liquid chromatography (HPLC). Electrospray ionization (ESI) tandem mass spectrometry of the eluted peptides yielded peptide mass spectra and fragment ion mass spectra. Comparison with databases identified the respective *Salmonella* proteins.

production yielded 219 candidates, including 152 previously characterized enzymes (90 essential, 62 contributing or dispensable) and 67 new candidates. Twenty-nine of these candidates (Supplementary Table 8) have been reported to be essential *in vitro* in *E. coli*, suggesting that they might be also essential in *Salmonella* during infection. Validation of these genes would require conditional mutants expressing the candidate enzyme in *in vitro* culture but not during infection. Unfortunately, suitable expression cassettes with tight *in vivo* repression¹⁹ have not yet been described.

The remaining 38 new and testable candidates belonged to five group A biosynthesis pathways (biosynthesis of riboflavin, ubiquinone, menaquinone, unsaturated fatty acids and diaminopimelic acid). We determined their *in vivo* relevance by inactivating key enzymes in each pathway. All pathway defects except for inactive menaquinone biosynthesis markedly reduced *Salmonella* growth in the mouse typhoid fever model, supporting the utility of our category-based enzyme selection strategy (Supplementary Table 6; results for *asd* were compatible with indirect observations using vaccine constructs²³, and results for *ribB* were compatible with our recent findings²⁴).

On the other hand, the rather limited number of new testable candidates was initially disappointing and suggested that our restriction to only a few pathway categories might have been too stringent. We therefore inactivated a further 21 enzymes from other pathway categories (involved in nutrient uptake and catabolism, group B biosynthesis and redundant functions). We selected these enzymes from the *in vivo* *Salmonella* proteome data on the basis of the strong association observed between detectable enzyme expression and functional relevance (see above). Of the 21 mutants, 16 had decreased growth rates in the typhoid fever model (Supplementary Table 6), supporting the utility of our *in vivo* proteome data for guiding the search for new and relevant metabolic processes. On the other hand, all 21 mutants were still quite virulent (Supplementary Table 6), supporting our hypothesis that redundant enzymes and enzymes involved in nutrient transport, catabolism or group B biosynthesis are mostly non-essential. Therefore, our initial genome-wide search for essential enzymes in the biosynthesis group A, central intermediary metabolism and energy production categories might have been nearly exhaustive. These new mutant phenotypes complement our previous analysis, providing combined information regarding 712 metabolic *Salmonella* enzymes (126 essential, 126 contributing, 282

dispensable and 178 detected but not yet functionally characterized; Supplementary Table 3) and at least 278 metabolites (Supplementary Table 7) in the typhoid fever model.

In the enteritis model, we tested 23 enzymes that were selected from the enteritis proteome data (Supplementary Table 5). Owing to differences between the typhoid fever and enteritis proteome data sets, this included some new enzymes (Supplementary Table 6). All identified essential enzymes belonged to biosynthesis group A, whereas enzymes involved in nutrient transport, catabolism or group B biosynthesis were all non-essential. Combined, our data provides information regarding 744 enzymes and 369 different metabolites in the enteritis model (Supplementary Tables 3 and 7).

Salmonella nutrition during infection

Only a minority of metabolic enzymes are absolutely required for systemic *Salmonella* virulence (126 out of 534 functionally characterized enzymes). Owing to biased interest in functionally relevant enzymes in this and previous studies, the actual percentage of essential metabolic enzymes might be even lower. This remarkable robustness (resilience against disturbances) is due in part to extensive network redundancy (Fig. 3c), as previously observed for other metabolic networks²⁵.

In addition, environmental factors could influence which enzymes are essential. *Salmonella* has been assumed to reside in nutrient-poor macrophage phagosomes during systemic infection^{26,27}. Indeed, several group A metabolites were apparently not available at sufficiently high concentrations (in the range of 100 μ M) to supplement the respective *Salmonella* biosynthetic mutants. On the other hand, all characterized biosynthesis group B mutants retained substantial virulence in the typhoid fever model (Supplementary Tables 1, 2, 3 and 6), but these auxotrophs could not grow without supplementation, indicating that various amino acids, purine and pyrimidine nucleosides, and vitamins were provided by the host in significant quantities (Supplementary Table 9). The partial attenuation of some catabolic and transporter mutants suggests that *Salmonella* utilizes host nucleosides and RNA, glycerol, sialic acid, hexoses and pentoses, vitamins and both aerobic and anaerobic electron acceptors during typhoid fever (Supplementary Tables 3 and 9). The availability of apparently diverse nutrients in the host might contribute substantially to the robustness of *Salmonella* metabolism during typhoid fever.

Shortage of new drug targets

Most of the characterized *Salmonella* metabolic enzymes are unsuitable as antimicrobial targets, as their inactivation failed to abolish *Salmonella* virulence. Of the 155 more promising target candidates (126 essential enzymes and 29 predicted essential enzymes; Supplementary Tables 3 and 8), 64 are conserved in a diverse set of major human pathogens² (*Staphylococcus aureus*, *Enterococcus faecalis*, *Streptococcus pneumoniae* and *Haemophilus influenzae*; Supplementary Table 8), and could thus represent attractive broad-spectrum antimicrobial targets. However, almost all of these targets belong to pathways already inhibited by current antibiotics (tRNA synthases, peptidoglycan biosynthesis, folate biosynthesis, isoprenoid biosynthesis, fatty acid biosynthesis) or pathways previously considered for antimicrobial development^{28–31}. Eight newly identified candidates all have very high sequence identities to human enzymes with key roles in central metabolism (blastp expectation values between 10^{-45} and 10^{-124} ; Supplementary Table 8), suggesting a direct risk for unwanted side effects^{2,28,29}. Collectively, these data suggest a shortage of new targets for broad-spectrum antimicrobial chemotherapy.

Discussion

Metabolic enzymes could represent attractive new targets for the development of urgently needed antibiotics. However, our systematic network analysis of *Salmonella* metabolism during infection

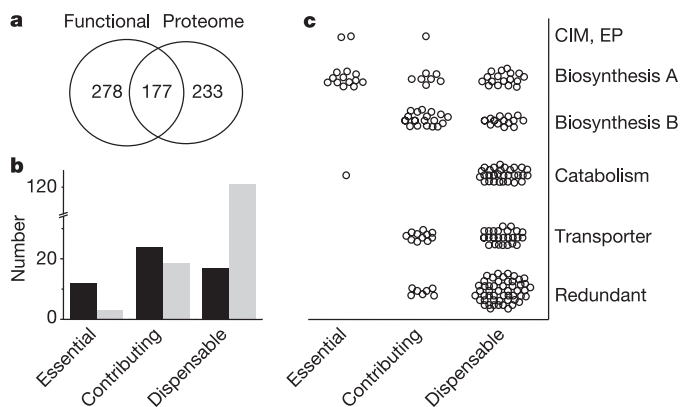


Figure 3 | Properties of *Salmonella* metabolic enzymes during typhoid fever. **a**, Summary of functional information and proteome data. **b**, Functionally important enzymes are over-represented among proteins detected by *in vivo* proteomics (black bars, enzyme belongs to detected pathway; grey bars, not detected; χ^2 test, $n = 196$, $P < 0.0001$). **c**, Essential enzymes belong to a small subgroup of pathways (CIM, central intermediary metabolism; EP, energy production; Biosynthesis A, biosynthesis of products that *Salmonella* cannot efficiently obtain from the host; Biosynthesis B, biosynthesis of metabolites that *Salmonella* can readily obtain and use from external sources; χ^2 test, $n = 196$; $P < 0.0001$).

has revealed that the large majority of *Salmonella* enzymes are non-essential for *Salmonella* virulence. This remarkable robustness is largely due to a combination of metabolic network redundancy and apparently nutrient-rich host environments that render *Salmonella* partially independent of many biosynthetic and catabolic capabilities. Absolutely essential enzymes are almost exclusively found in pathways relating to the biosynthesis of compounds that *Salmonella* cannot efficiently obtain from the host, or in central intermediary metabolism and energy production. Unfortunately, almost all of the 155 essential enzymes we identified are either absent in other major pathogens or represent already known antimicrobial targets.

Owing to incomplete experimental data, it is likely that some attractive metabolic targets remain unidentified. On the other hand, the observed very low hit frequency (no promising new target among 534 characterized enzymes) suggests that the total number of new and attractive metabolic targets for broad-spectrum antimicrobial development is probably limited. This could in part explain why so few new classes of antibiotics have been approved in the past 30 years¹. However, our finding that several still-unexploited 'old' enzymes might be the most attractive targets for antimicrobial chemotherapy⁸ could accelerate the development of urgently needed new antibiotics.

METHODS

Salmonella strains, molecular biology techniques, generation of *Salmonella* metabolic network overviews, mouse infection and statistical methods are described in the Supplementary Methods.

Flow cytometric purification of *Salmonella* from infected mice. Mice were infected with *Salmonella enterica* serovar Typhimurium expressing GFP at moderate, non-attenuating levels³². Fluorescent GFP-expressing *Salmonella* were sorted from spleen homogenates (typhoid fever model¹²) or caecum homogenates (enteritis model¹¹) using two-colour flow cytometry³³ (see Supplementary Methods). Cytometric *Salmonella* counts and colony-forming unit numbers in the sorted sample matched within $\pm 20\%$, indicating that almost all sorted *Salmonella* were alive. Control experiments revealed that *Salmonella* protein degradation and *de novo* biosynthesis were efficiently blocked during sorting (see Supplementary Methods), suggesting largely unchanged *in vivo* enzyme composition in the purified *Salmonella*.

Protein identification using mass spectrometry. Sorted *Salmonella* were digested with trypsin in urea-containing buffer. The resulting peptide mixtures were analysed by nanoscale liquid chromatography–tandem mass spectrometry (LC–MS/MS) on an ion trap Fourier-transform mass spectrometer capable of very high mass accuracy and sequencing speed (Thermo Electron, LTQ-FT; see Supplementary Methods). Tandem mass spectra were searched in the non-redundant NCBI sequence database, considering only peptides conforming to full tryptic specificity and for which the molecular mass matched the calculated mass within 10 p.p.m. (see Supplementary Methods). These peptides were used to identify *Salmonella* proteins with high stringency. Using an inverted 'decoy' database we estimated that at most one or two protein identifications might be false-positive (see Supplementary Methods).

Received 12 December 2005; accepted 1 February 2006.

- Norrby, S. R., Nord, C. E. & Finch, R. Lack of development of new antimicrobial drugs: a potential serious threat to public health. *Lancet Infect. Dis.* **5**, 115–119 (2005).
- Black, M. T. & Hodgson, J. Novel target sites in bacteria for overcoming antibiotic resistance. *Adv. Drug Deliv. Rev.* **57**, 1528–1538 (2005).
- Hughes, D. Exploiting genomics, genetics and chemistry to combat antibiotic resistance. *Nature Rev. Genet.* **4**, 432–441 (2003).
- Guimera, R. & Nunes Amaral, L. A. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
- Price, N. D., Reed, J. L. & Pálsson, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Rev. Microbiol.* **2**, 886–897 (2004).
- Badarinarayana, V. et al. Selection analyses of insertional mutants using subgenic-resolution arrays. *Nature Biotechnol.* **19**, 1060–1065 (2001).
- Kuepfer, L., Sauer, U. & Blank, L. M. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res.* **15**, 1421–1430 (2005).
- Thomson, C. J., Power, E., Ruebsamen-Waigmann, H. & Labischinski, H.

- Antibacterial research and development in the 21st Century—an industry perspective of the challenges. *Curr. Opin. Microbiol.* **7**, 445–450 (2004).
- Bhan, M. K., Bahl, R. & Bhatnagar, S. Typhoid and paratyphoid fever. *Lancet* **366**, 749–762 (2005).
- O'Ryan, M., Prado, V. & Pickering, L. K. A millennium update on pediatric diarrheal illness in the developing world. *Semin. Pediatr. Infect. Dis.* **16**, 125–136 (2005).
- Hapfelmeier, S. & Hardt, W. D. A mouse model for *S. typhimurium*-induced enterocolitis. *Trends Microbiol.* **13**, 497–503 (2005).
- Santos, R. L. et al. Animal models of *Salmonella* infections: enteritis versus typhoid fever. *Microbes Infect.* **3**, 1335–1344 (2001).
- McClelland, M. et al. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature* **413**, 852–856 (2001).
- Porwollik, S., Santiviago, C. A., Cheng, P., Florea, L. & McClelland, M. Differences in gene content between *Salmonella enterica* serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin. *J. Bacteriol.* **187**, 6545–6555 (2005).
- Deng, W. et al. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J. Bacteriol.* **185**, 2330–2337 (2003).
- Parkhill, J. et al. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413**, 848–852 (2001).
- McClelland, M. et al. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nature Genet.* **36**, 1268–1274 (2004).
- Keseler, I. M. et al. EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* **33**, D334–D337 (2005).
- Knuth, K., Niesalla, H., Hueck, C. J. & Fuchs, T. M. Large-scale identification of essential *Salmonella* genes by trapping lethal insertions. *Mol. Microbiol.* **51**, 1729–1744 (2004).
- Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
- Giaever, G. et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
- Baumler, A. J., Kusters, J. G., Stojiljkovic, I. & Heffron, F. *Salmonella typhimurium* loci involved in survival within macrophages. *Infect. Immun.* **62**, 1623–1630 (1994).
- Curtiss, R. III, Nakayama, K. & Kelly, S. M. Recombinant avirulent *Salmonella* vaccine strains with stable maintenance and high level expression of cloned genes *in vivo*. *Immunol. Invest.* **18**, 583–596 (1989).
- Rollenhagen, C. & Bumann, D. *Salmonella enterica* highly expressed genes are disease-specific. *Infect. Immun.* **74**, 1649–1660 (2006).
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
- Eriksson, S., Lucchini, S., Thompson, A., Rhen, M. & Hinton, J. C. Unravelling the biology of macrophage infection by gene expression profiling of intracellular *Salmonella enterica*. *Mol. Microbiol.* **47**, 103–118 (2003).
- Henry, T., Garcia-del Portillo, F. & Gorvel, J. P. Identification of *Salmonella* functions critical for bacterial cell division within eukaryotic cells. *Mol. Microbiol.* **56**, 252–267 (2005).
- Freiberg, C. et al. Identification of novel essential *Escherichia coli* genes conserved among pathogenic bacteria. *J. Mol. Microbiol. Biotechnol.* **3**, 483–489 (2001).
- Gerdes, S. Y. et al. From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways. *J. Bacteriol.* **184**, 4555–4572 (2002).
- Heath, R. J., White, S. W. & Rock, C. O. Lipid biosynthesis as a target for antibacterial agents. *Prog. Lipid Res.* **40**, 467–497 (2001).
- Andries, K. et al. A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* **307**, 223–227 (2005).
- Wendland, M. & Bumann, D. Optimization of GFP levels for analyzing *Salmonella* gene expression during an infection. *FEBS Lett.* **521**, 105–108 (2002).
- Bumann, D. Examination of *Salmonella* gene expression in an infected mammalian host using the green fluorescent protein and two-colour flow cytometry. *Mol. Microbiol.* **43**, 1269–1283 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Förster for support, T. Aebischer for discussion, M. Sörensen, K. Raba and C. Reimer for technical assistance, and P. Mortensen and J. V. Olsen for providing a script to calibrate the LTQ-FT data. This work was supported by grants from the Deutsche Forschungsgemeinschaft (to D. Bumann).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D. Bumann (bumann.dirk@mh-hannover.de).

LETTERS

A magnetic torsional wave near the Galactic Centre traced by a 'double helix' nebula

Mark Morris¹, Keven Uchida² & Tuan Do¹

The magnetic field in the central few hundred parsecs of the Milky Way has a dipolar geometry and is substantially stronger than elsewhere in the Galaxy, with estimates ranging up to a milligauss (refs 1–6). Characterization of the magnetic field at the Galactic Centre is important because it can affect the orbits of molecular clouds by exerting a drag on them, inhibit star formation, and could guide a wind of hot gas or cosmic rays away from the central region. Here we report observations of an infrared nebula having the morphology of an intertwined double helix about 100 parsecs from the Galaxy's dynamical centre, with its axis oriented perpendicular to the Galactic plane. The observed segment is about 25 parsecs in length, and contains about 1.25 full turns of each of the two continuous, helically wound strands. We interpret this feature as a torsional Alfvén wave propagating vertically away from the Galactic disk, driven by rotation of the magnetized circumnuclear gas disk. The direct connection between the circumnuclear disk and the double helix is ambiguous, but the images show a possible meandering channel that warrants further investigation.

The new image was obtained using the Multiband Imaging Photometer (MIPS) camera⁷ on the Spitzer Space Telescope at a wavelength of 24 μm . It reveals the elongated, double helix nebula (DHN) shown in Fig. 1. This nebula clearly extends beyond the edge of the observed field. The full extent of the DHN is evident in mid-infrared images previously obtained with the Midcourse Space Experiment (MSX) satellite (Fig. 10 of ref. 8; see also ref. 26). In the MSX data (Fig. 2), it is possible with hindsight to recognize the double helix, although the sensitivity and spatial resolution are not sufficient to have revealed this structure before the Spitzer observation. The MSX images show that the structure extends towards smaller Galactic latitude (b) from the region observed with Spitzer; it is at least 20 arcmin (50 pc) in length, extending between Galactic longitude $l = 0.08$, $b = 0.5$ and $l = 0.02$, $b = 0.80$, and it is possibly part of a larger structure (see below). However, there is no evidence in the MSX data that the helical strands persist outside the region observed with MIPS. It is also noteworthy that there is no sign in MSX images of a negative-latitude counterpart to the DHN.

The four wavelengths observed with the MSX satellite (8.3, 12.1, 14.7 and 21.3 μm) provide a sampling of the spectral energy distribution of the mid-infrared emission. The surface brightnesses at a few sample locations, measured relative to the local background, uniformly indicate that the emission is substantially strongest at the shorter wavelengths (detailed in Supplementary Information). The best-fit colour temperature is 630 ± 40 K. If the emission corresponds to thermal emission from dust for which the mid-infrared emissivity is proportional to ν^p , where ν is the frequency, then the implied best-fit dust temperature is 410 ± 16 K ($p = 1$) or 310 ± 10 K ($p = 2$). If the two shortest-wavelength channels of MSX are affected by strong emission from bands of polycyclic aromatic hydrocarbon molecules (PAHs), as is frequently the case,

then these temperatures might be strongly overestimated, and detailed spectroscopy will be needed to infer the physical dust temperature.

If the high dust temperatures are upheld, two possible dust heating mechanisms that might be considered are: (1) heating by streaming of gas particles with respect to the dust (for example, as in a Galactic wind⁹) and (2) impulsive heating of small grains by ambient Galactic starlight. For reasonable densities of the streaming gas, the former mechanism provides inadequate heating, even for a relative gas–dust drift velocity as large as the Alfvén velocity in this region ($\sim 10^3 \text{ km s}^{-1}$, see below). The second mechanism warrants further investigation. Ignoring dust extinction, the equilibrium temperature of dust in the Galactic bulge, 100 pc from the Galactic Centre, is only 30–40 K, but small grains can undergo large temperature excursions as they absorb ultraviolet photons¹⁰, and the total mass required in small grains to account for the DHN emission, although sensitive to the assumed impinging ultraviolet intensity and grain size distribution, is much less than a solar mass.

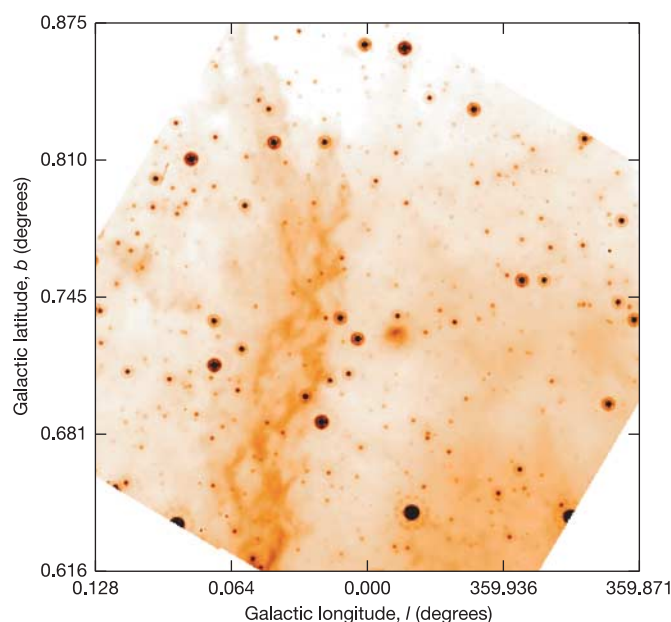


Figure 1 | The double helix nebula (DHN), observed at the infrared wavelength of 24 μm with the MIPS camera on the Spitzer Space Telescope. The spatial resolution is 6 arcsec. At the 8 kpc distance of the Galactic Centre, 1 arcmin corresponds to 2.5 pc. The full region observed extends well to the lower right of the region shown, and consists of a long strip centred on the bright infrared source AFGL5376²⁵, upon which we will report separately.

¹Department of Physics and Astronomy, University of California, Los Angeles, California 90095-1547, USA. ²Center for Radiophysics and Space Research, Cornell University, Space Sciences Building, Ithaca, New York 14853-6801, USA.

Two potential alternatives to thermal dust emission—thermal bremsstrahlung emission by hot electrons and non-thermal emission from a population of relativistic electrons—are both rendered unlikely by the observed spectral slope, as detailed in Supplementary Information. In any case, the average variance of the data from the best-fitting non-thermal power-law models is twice as large as that of the best-fitting thermal model—that with $p = 2$. Consequently, we conclude that the emission observed from the double helix most probably arises from thermal dust emission. Shorter-wavelength observations in the four bands of the Infrared Array Camera (IRAC) on Spitzer will help clarify this point.

The sky location and orientation of the DHN are very suggestive of an association with the centre of the Galaxy. Not only does the long axis of this structure point roughly to the Galactic Centre, less than a degree away, but it is also oriented along the Galaxy's axis of rotation. Furthermore, as we argue below, an association with the Galactic Centre provides a natural explanation for this structure. At an assumed Galactic Centre distance of 8 kpc, the wavelength of the individual strands of the double helix—about 7.5 arcmin—corresponds to a length of 19 pc, and the maximum strand separation (that is, the overall width of the structure) is 1.4 arcmin, or ~ 3.5 pc. The individual strands are just barely resolved, with a width of about 7 or 8 arcsec at their narrowest locations, compared to the full-width at half-maximum of the MIPS point spread function at 24 μm , which is 6 arcsec.

The observed helical structure is far too large to be attributed to stellar activity. Rather, it probably results from a dynamically ordered, large-scale, interstellar phenomenon involving interstellar gas, dust and magnetic fields. We propose that the DHN is a magnetohydrodynamic torsional Alfvén wave propagating more or less vertically out of the Galactic plane, along magnetic field lines, from the near vicinity of the Galactic Centre. This hypothesis conforms to the apparent, global dipolar geometry of the Galactic Centre magnetic field¹. It is natural to ascribe the driving of the torsional wave to rotation about the Galactic Centre, and an obvious candidate to do this is the circumnuclear disk (CND)^{1,11–14}. The characteristics of the CND match those needed to produce the characteristics of the

proposed torsional wave: it has an inner radius of ~ 1 pc, and extends out to several parsecs, being somewhat asymmetric in its outer regions, possibly because of an interaction with the SgrA East supernova remnant^{1,15–17}. The lateral extent of the proposed torsional wave, ~ 3.5 pc, is consistent with the planar extent of the CND, 2–7 pc, and thus with the hypothesis that the rotation of the CND is responsible for a torsional wave propagating through a uniform field. Little lateral growth of the helical structure is expected as long as the Alfvén speed is much greater than the rotation speed of the driving disk, or as long as the longitudinal wavelength is much greater than the lateral extent of the feature, which is clearly the case. Furthermore, the CND is strongly magnetized, and its predominant shear-induced azimuthal field is believed to merge smoothly with the ambient vertical Galactic Centre field^{18–20}.

The rotation velocity of the CND is approximately constant at 100 km s^{-1} , giving a period of $10^4 R$ years at radius R (where R is in units of pc). Drawing a correspondence between the period of the CND and the wavelength of the double helix, we can derive the Alfvén velocity: $V_A = 10^3 \text{ km s}^{-1}$. This, in turn, can be used to estimate the magnetic field strength, B , in this region, with $V_A^2 = B^2 / (4\pi m_p n_p)$, where n_p is the proton density in the medium through which the wave propagates, and m_p is the proton mass. This gives $B = 0.5 n_p^{1/2} \text{ mG}$. The density in this region is not known; if the medium through which the wave propagates is the hot (10^8 K) medium evidenced in diffuse X-rays²¹, then $n_p \approx 0.1 \text{ cm}^{-3}$, and consequently, $B = 0.1 \text{ mG}$. The DHN is located on the outskirts of the region to which the 10^8 K gas may extend²², but given previous estimates for B ranging from 0.01 to 1 mG (refs 1, 4, 5), this is a plausible field strength. If, on the other hand, the field strength is that estimated from the rigidity of the non-thermal radio filaments, $B \approx 1 \text{ mG}$ (ref. 1), we infer a local density of $n_p \approx 5 \text{ cm}^{-3}$, which does not violate any observational constraints.

A possible weakness of this hypothesis is that the torsional wave cannot yet be followed all the way down to its hypothetical source, the CND, presumably because of the enhanced confusion by intervening material and superimposed emission structures closer to the Galactic plane. The MSX data do, however, show a potential meandering channel along which the wave might propagate (Fig. 2). Such a meander can be ascribed to the kink instability, arising naturally in a twisted magnetic field. Although the contrast of this potential channel with the local background emission is evidently weak, there are two locations at which evidence for a channel exists in the form of parallel, linear emission features having a separation (~ 5 pc) comparable to the width of the helical structure (marked with arrows in Fig. 2). These features (shown in greater detail in Supplementary Information) can be interpreted as limb-brightened cylinders having relatively thin, emitting walls. Patchy absorption abutting the brighter, lower-latitude one of these indicates the presence of a nearby concentration of dust and gas, so the emission from this apparent part of the channel might be attributable to an interaction between the magnetic energy in the torsional wave channel and the surrounding, relatively dense, interstellar material. (The recently released images from IRAC, however, indicate that this patchy absorption is likely to be in the relatively near foreground (S. Stolovy, personal communication).) One might also speculate that the meander of the apparent channel is partially caused by this interaction, deflecting the wave energy towards positive Galactic longitude. The absence of a negative-latitude counterpart is another potential weakness of the torsional wave hypothesis, inasmuch as such waves should propagate equally in both directions away from the driving disk, if that disk is symmetric about its midplane. However, the MSX images show a brighter, more complex background at negative latitudes, so some combination of background confusion and dissipative shock interactions—both common in the Galactic Centre—could account for the absence of a counterpart.

One question that our hypothesis leaves unanswered is why the helical structure has two strands. A uniform, axisymmetric, rotating

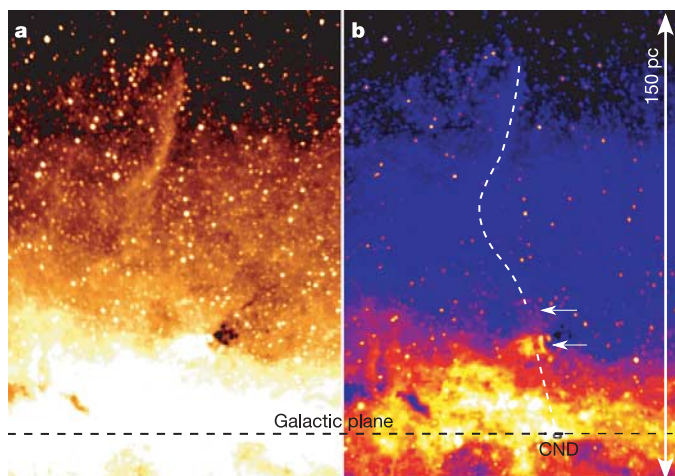


Figure 2 | Images of a 1° field (from -0.1° to $+0.9^\circ$ in Galactic latitude) that include the Galactic plane, the Galactic Centre, and the DHN. These data were taken with the MSX satellite. The spatial resolution of these images is 20 arcsec. **a**, A-band image ($8.3 \mu\text{m}$) with a colour scale chosen to emphasize the DHN. **b**, Same image with a different colour scale. The relative locations and sizes of the DHN, the circumnuclear disk (CND), and the proposed channel linking them, are all shown, along with the two bright channel segments marked with arrows and described in the text. These channel segments are located at Galactic coordinates (l, b) $-0.01^\circ, 0.15^\circ$ and $0.0^\circ, 0.23^\circ$. Additional MSX images of the channel segments are available in Supplementary Information.

disk driving a torsional wave in a field perpendicular to the disk would produce a cylindrically symmetric structure. The presence of two strands indicates that the driver has an $m = 2$ symmetry (surface density has a term of the form $\exp(im\phi)$, where ϕ is the azimuthal angle). This could take the form of a bar, or, in the extreme case, one could attribute the strands of the double helix to two diametrically opposed blobs into which the vertical magnetic flux threading the disk had been concentrated. As the timescale for propagation of an Alfvén wave from the CND to the observed double helix, ~ 100 pc away, is 10^5 yr, or $10/R$ rotation periods of the disk, there has been sufficient time for the strong shear in the CND to have eliminated any $m = 2$ deviation from axisymmetry that may have been present when the double helix was launched. Nonetheless, the CND is somewhat non-axisymmetric at present, and its inner portions have two prominent concentrations of material on opposite sides of the centre along the major axis of the projected disk^{11,13,14,23,24}. Consequently, it is possible that the $m = 2$ deviation from axisymmetry is still in place, or that it has re-formed since the double helix was launched.

Finally, if the favoured mechanism for the mid-infrared emission from the DHN remains thermal dust emission, then we must face the question of why dust is present at all in the torsional wave. If the emission is from small dust grains, then such grains are likely to carry a net charge, in which case they can then be carried aloft by the torsional wave.

Received 23 October; accepted 15 December 2005.

- Morris, M. & Serabyn, E. The Galactic center environment. *Annu. Rev. Astron. Astrophys.* **34**, 645–701 (1996).
- Yusef-Zadeh, F., Morris, M. & Chance, D. Large, highly organized radio structures near the galactic centre. *Nature* **310**, 557–561 (1984).
- Lang, C. C., Morris, M. R. & Echevarria, L. A radio polarimetric study of the Galactic center threads. *Astrophys. J.* **526**, 727–743 (1999).
- Yusef-Zadeh, F. The origin of the Galactic center nonthermal radio filaments: young stellar clusters. *Astrophys. J.* **598**, 325–333 (2003).
- Yusef-Zadeh, F., Hewitt, J. W. & Cotton, W. A 20 centimeter survey of the galactic center region. I. detection of numerous linear filaments. *Astrophys. J.* **155** (Suppl.), 421–550 (2004).
- LaRosa, T. N. et al. Evidence of a weak Galactic center magnetic field from diffuse low-frequency nonthermal radio emission. *Astrophys. J.* **626**, L23–L27 (2005).
- Rieke, G. H. et al. The Multiband Imaging Photometer for Spitzer (MIPS). *Astrophys. J.* **154** (Suppl.), 25–29 (2004).
- Price, S. D., Egan, M. P., Carey, S. J., Mizuno, D. R. & Kuchar, T. A. Midcourse Space Experiment survey of the Galactic plane. *Astron. J.* **121**, 2819–2842 (2001).
- Belmont, R., Tagger, M., Muno, M., Morris, M. & Cowley, S. A hot helium plasma in the Galactic center region. *Astrophys. J.* **631**, L53–L56 (2005).
- Puget, J. L. & Léger, A. A new component of the interstellar matter - small grains and large aromatic molecules. *Annu. Rev. Astron. Astrophys.* **27**, 161–198 (1989).
- Güsten, R. et al. Aperture synthesis observations of the circumnuclear ring in the Galactic center. *Astrophys. J.* **318**, 124–138 (1987).
- Latvakoski, H. M., Stacey, G. J., Gull, G. E. & Hayward, T. L. Kuiper Widefield Infrared Camera far-infrared imaging of the galactic center: the circumnuclear disk revealed. *Astrophys. J.* **511**, 761–773 (1999).
- Yusef-Zadeh, F., Stolovy, S. R., Burton, M., Wardle, M. & Ashley, M. C. B. High spectral and spatial resolution observations of shocked molecular hydrogen at the Galactic center. *Astrophys. J.* **560**, 749–762 (2001).
- Christopher, M. H., Scoville, N. Z., Stolovy, S. R. & Yun, M. S. HCN and HCO⁺ observations of the Galactic circumnuclear disk. *Astrophys. J.* **622**, 345–365 (2005).
- Yusef-Zadeh, F. et al. in *The Central Parsecs of the Galaxy* (eds Falcke, H., Cotera, A., Duschl, W. J., Melia, F. & Rieke, M. J.) 197–213 (ASP Conf. Ser. No. 186, ASP, San Francisco, 1999).
- Maeda, Y. et al. A Chandra study of Sagittarius A East: a supernova remnant regulating the activity of our Galactic center? *Astrophys. J.* **570**, 671–687 (2002).
- Rockefeller, G., Fryer, C. L., Baganoff, F. K. & Melia, F. The x-ray ridge surrounding Sgr A* at the Galactic center. *Astrophys. J.* **635**, L141–L144 (2005).
- Hildebrand, R. H. et al. Polarization of the thermal emission from the dust ring at the center of the Galaxy. *Astrophys. J.* **417**, 565–571 (1993).
- Wardle, M. & Königl, A. A model for the magnetic field in the molecular disk at the Galactic center. *Astrophys. J.* **362**, 120–134 (1990).
- Novak, G. et al. Submillimeter polarimetric observations of the Galactic center. *Astrophys. J.* **529**, 241–250 (2000).
- Muno, M. P. et al. Diffuse x-ray emission in a deep Chandra image of the Galactic center. *Astrophys. J.* **613**, 326–342 (2004).
- Yamauchi, S., Kawada, M., Koyama, K., Kunieda, H. & Tawara, Y. Optically thin hot plasma near the Galactic center - Mapping observations of the 6.7 keV iron line. *Astrophys. J.* **365**, 532–538 (1990).
- Genzel, R., Watson, D. M., Crawford, M. K. & Townes, C. H. The neutral-gas disk around the galactic center. *Astrophys. J.* **297**, 766–786 (1985).
- Gatley, I. et al. The spatial distribution and velocity field of the molecular hydrogen line emission from the centre of the Galaxy. *Mon. Not. R. Astron. Soc.* **222**, 299–306 (1986).
- Uchida, K. I., Morris, M. R., Serabyn, E. & Bally, J. AFGL 5376: a strong, large-scale shock near the Galactic center. *Astrophys. J.* **421**, 505–516 (1994).
- Bland-Hawthorn, J. & Cohen, M. The large-scale bipolar wind in the Galactic center. *Astrophys. J.* **582**, 246–256 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements M.M. is grateful to S. Cowley and S. Stolovy for discussions. This work is based on observations made with the Spitzer Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. Partial support for this work was provided by NASA through an award issued by JPL/Caltech. This research also made use of data products from the Midcourse Space Experiment, available via the NASA/IPAC Infrared Science Archive.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.M. (morris@astro.ucla.edu).

Discovery of two young brown dwarfs in an eclipsing binary system

Keivan G. Stassun¹, Robert D. Mathieu² & Jeff A. Valenti³

Brown dwarfs are considered to be ‘failed stars’ in the sense that they are born with masses between the least massive stars ($0.072 M_{\odot}$)¹ and the most massive planets ($\sim 0.013 M_{\odot}$)²; they therefore serve as a critical link in our understanding of the formation of both stars and planets³. Even the most fundamental physical properties of brown dwarfs remain, however, largely unconstrained by direct measurement. Here we report the discovery of a brown-dwarf eclipsing binary system, in the Orion Nebula star-forming region, from which we obtain direct measurements of mass and radius for these newly formed brown dwarfs. Our mass measurements establish both objects as brown dwarfs, with masses of $0.054 \pm 0.005 M_{\odot}$ and $0.034 \pm 0.003 M_{\odot}$. At the same time, with radii relative to the Sun’s of $0.669 \pm 0.034 R_{\odot}$ and $0.511 \pm 0.026 R_{\odot}$, these brown dwarfs are more akin to low-mass stars in size. Such large radii are generally consistent with theoretical predictions for young brown dwarfs in the earliest stages of gravitational contraction^{4,5}. Surprisingly, however, we find that the less-massive brown dwarf

is the hotter of the pair; this result is contrary to the predictions of all current theoretical models of coeval brown dwarfs.

Mass is the most fundamental property of a brown dwarf, as it determines all other physical characteristics and governs how brown dwarfs evolve with time. Theoretical brown-dwarf evolution models^{2,4,5} make specific, testable predictions of the relationship between a brown dwarf’s mass and its other physical properties, such as radius and temperature. Unfortunately, these predictions have yet to be empirically tested because the basic physical properties of brown dwarfs are extremely difficult to measure; the mass of only one brown dwarf has been measured with sufficient accuracy to make the brown-dwarf nature of the object definitive⁶, and in no case has a brown dwarf’s radius been measured directly. Astronomers have traditionally used stellar eclipsing binaries to measure the masses and radii of stars⁷, but until now no brown-dwarf eclipsing binaries were known. Here we have discovered the 2-Micron All-Sky Survey (2MASS) object J05352184–0546085 to be a previously unknown eclipsing binary in the Orion Nebula, and the first example of an

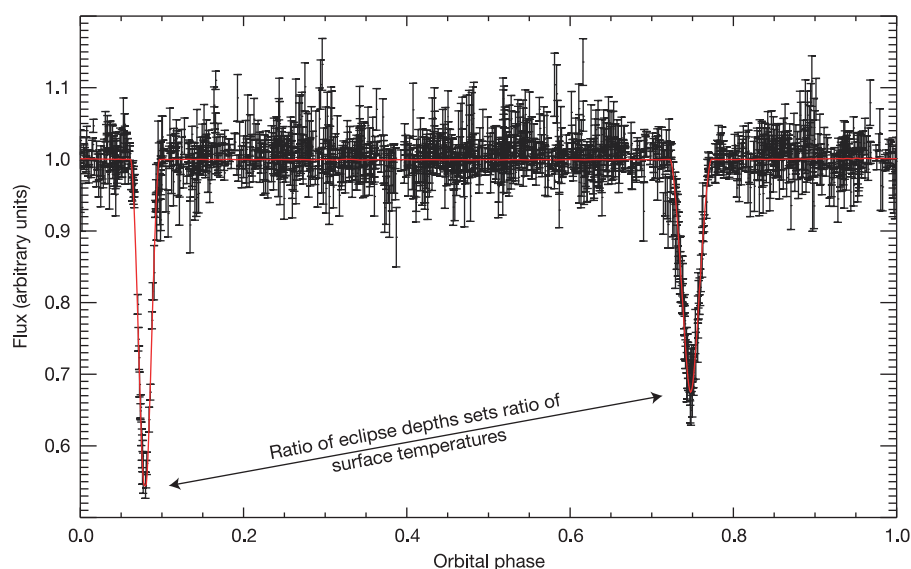


Figure 1 | Light curve of 2MASS J05352184–0546085 at $0.8 \mu\text{m}$. We repeatedly imaged 2MASS J05352184–0546085 with the 0.9-m telescope at Kitt Peak National Observatory and with the 0.9-m, 1.0-m and 1.3-m telescopes at the Cerro Tololo Inter-American Observatory, from December 1994 to March 2005. In total, 1,590 flux measurements were obtained on 280 separate nights and with an average cadence of 5–6 measurements per night. Error bars show ± 1 s.d.; the typical uncertainty is 2%. A time-series analysis reveals an unambiguous period of $P = 9.779621 \pm 0.000042$ days. The data are shown here folded on this period and phased relative to

periastron passage, as determined from the orbit solution (Fig. 2). The ratio of eclipse depths provides a direct measure of the ratio of surface temperatures, with the deeper eclipse corresponding to the eclipse of the hotter component. In this system, the deeper eclipse corresponds to the eclipse of the lower-mass component (see Fig. 2). A model fit to the light curve incorporating the final orbital and physical parameters of 2MASS J05352184–0546085 (Table 1, column 2) is also shown (solid red curve). The average residual relative to this model is 1.9%, and the reduced χ -squared of the fit is 0.85.

¹Department of Physics and Astronomy, Vanderbilt University, Nashville, Tennessee 37235, USA. ²Department of Astronomy, University of Wisconsin—Madison, Madison, Wisconsin 53706, USA. ³Space Telescope Science Institute, Baltimore, Maryland 21218, USA.

eclipsing binary comprising two brown dwarfs (a summary diagram of the salient features of 2MASS J05352184–0546085 is provided in the Supplementary Information).

The light curve of 2MASS J05352184–0546085 (Fig. 1) is characterized by two distinct, periodically recurring diminutions in flux that are well separated in phase, as is typical of eclipsing binaries in which the two components are fully detached (that is, not in contact). From radial velocity measurements of both components (Fig. 2) we obtain the binary mass ratio and derive a precise solution for all of the system parameters except the orbital inclination (Table 1, column 1). A combined analysis of the light curve and orbit solution^{8,9} then yields the inclination precisely and gives an accurate measurement of the total system mass (Table 1, column 2).

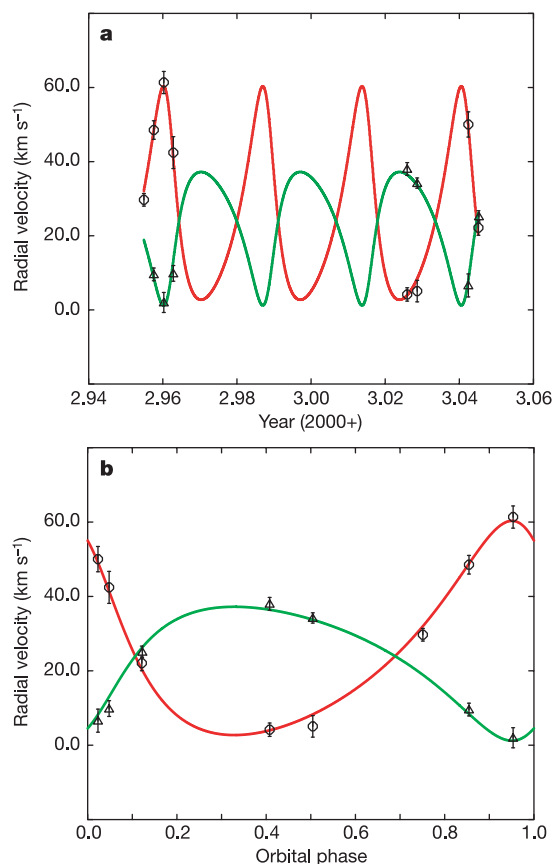


Figure 2 | Radial velocity measurements and orbit solution. We obtained high-resolution, near-infrared spectra of 2MASS J05352184–0546085 with the Phoenix spectrograph (resolving power of 30,000) at the 8-m Gemini South telescope on eight separate nights from December 2002 to January 2003. A cross-correlation analysis of the individual spectra against a radial-velocity standard star (see Fig. 3a) yields the radial velocities of the two components at eight distinct orbital phases. We tried several different radial-velocity standard stars with spectral types M1 to M9, and found that an M6.5 star produced the strongest cross-correlation signals for both components; the spectra of the two components are therefore probably similar to that of an M6.5 spectral type (see Fig. 3). Here the individual, heliocentric radial velocity measurements of the components of 2MASS J05352184–0546085 are shown, as functions of both time (a) and orbital phase (b). Orbital phase is measured with respect to the time of periastron passage (that is, closest approach of the two brown dwarfs to one another), as determined from the final orbit solution. Measurements of the primary are represented as triangles; the secondary as circles. Error bars show ± 1 s.d.; the typical uncertainty is 2 km s^{-1} . The final orbit solution (Table 1, column 2) was determined by simultaneously fitting the light-curve and radial-velocity measurements using a standard detached-eclipsing-binary model. This orbit solution is shown in a and b as solid curves, green for the more massive component, red for the less massive component. The reduced χ -squared of the fit is 1.0.

The individual masses follow directly from the total system mass and the binary mass ratio. At $M_1 = 0.0541 \pm 0.0046 M_\odot$ the more massive component—the ‘primary’—is four standard deviations below the $0.072 M_\odot$ threshold for a star, and at $M_2 = 0.0340 \pm 0.0027 M_\odot$ the lower-mass ‘secondary’ is even further below this threshold; both components are thus proved to be brown dwarfs. In addition, from the observed eclipse durations and orbital velocities we directly and accurately measure the radii of the brown dwarfs to be $R_1 = 0.669 \pm 0.034 R_\odot$ and $R_2 = 0.511 \pm 0.026 R_\odot$, representing the first direct measurements of brown-dwarf radii.

The relative depths of the eclipses (Fig. 1) yield the ratio of the brown dwarfs’ surface fluxes at $0.8 \mu\text{m}$, from which we determine the ratio of their surface temperatures to be $T_2/T_1 = 1.054 \pm 0.006$. Remarkably, the lower-mass secondary has a surface temperature that is slightly—but significantly—warmer than the primary. This follows directly from the fact that the deeper eclipse occurs when the secondary is eclipsed, which implies that it has the higher surface flux. Brown dwarf spectra deviate significantly from an ideal black-body, potentially affecting the eclipse depths. Examining up-to-date synthetic models of brown-dwarf spectra¹⁰ at temperatures appropriate to 2MASS J05352184–0546085 (Fig. 3), we find that the ratio of fluxes at $0.8 \mu\text{m}$ is consistent with that predicted from simple blackbodies to within 3%, which implies a ratio of temperatures that agrees with the blackbody value to within 0.5%, consistent with the uncertainty quoted above. Spots have been found on many brown dwarfs¹¹ and are likely to be present in 2MASS J05352184–0546085 as well. However, the effect of spots on the temperature ratio is likely to be minor^{12,13}.

The spectrum of the primary (Fig. 3a), as well as the observed near-infrared colours (Fig. 3b), both indicate a spectral type for the primary of $M6.5 \pm 0.5$ (ref. 14). This suggests a surface temperature of $T_1 = 2,650 \pm 100 \text{ K}$ (refs 12, 15) which, from the temperature ratio, then implies $T_2 = 2,790 \pm 105 \text{ K}$. The temperature scale of young brown dwarfs is uncertain, so the assignment of absolute temperatures to 2MASS J05352184–0546085 is subject to systematic uncertainties of $\sim 300 \text{ K}$ in addition to the formal spectral-typing errors quoted¹⁶. We emphasize that it is the ratio of surface temperatures that is most accurately determined from our analysis (Fig. 1).

The distance to 2MASS J05352184–0546085 can be determined by comparing its total luminosity with its observed apparent magnitude of $m_K = 13.58 \pm 0.02$ (ref. 17), and adopting a bolometric correction¹⁵ appropriate for the observed surface temperatures. The luminosity of each brown dwarf is calculated directly from its radius and surface temperature via the Stefan-Boltzmann relation, $L = 4\pi R^2 \sigma T^4$. The distance so determined is 435 ± 55 parsecs. Extinction by as much as 0.75 visual magnitudes may be present (Fig. 3), in which case the distance is 420 ± 55 parsecs. This distance is consistent with the distance to the Orion Nebula of 480 ± 80 parsecs (ref. 18). In addition, the centre-of-mass velocity (Table 1) is within 1 km s^{-1} of the systemic radial velocity ($25 \pm 1.5 \text{ km s}^{-1}$) of members in the Orion Nebula star-forming region^{13,19}. The Orion Nebula star cluster is extremely young, with an age that has been estimated to be just 1_{-1}^{+2} million years^{20,21}. 2MASS J05352184–0546085 is thus probably also very young, probably having formed within the past few million years. The near-infrared colours of 2MASS J05352184–0546085 (Fig. 3) limit the amount of remnant material available to the brown dwarfs for accretion. The observed masses are therefore unlikely to change significantly over time, and these brown dwarfs will probably forever remain brown dwarfs.

Models^{2,4,5} predict that very young brown dwarfs still in the early stages of gravitational contraction should be significantly larger and warmer than their more evolved counterparts, and that is in fact what we see in 2MASS J05352184–0546085. At an age of $\sim 1 \text{ Myr}$, these brown dwarfs are $\sim 500\%$ larger, and $\sim 1,500 \text{ K}$ warmer, than older brown dwarfs, which are predicted to have radii of $\sim 0.1 R_\odot$ and surface temperatures of $\sim 1,000 \text{ K}$ at 1 Gyr . The recently measured

Table 1 | Orbital and physical parameters of 2MASS J05352184–0546085

	Fit to radial-velocity data only	Fit to radial-velocity and light-curve data simultaneously
Time of periastron passage	2001.86334 ± 0.00024	2001.863650 ± 0.000095
Eccentricity, e	0.307 ± 0.024	0.3225 ± 0.0060
Orientation of periastron, ω	211.8 ± 3.9°	215.4 ± 1.1°
Semi-major axis, $a \sin i$	0.0401 ± 0.0015 AU	0.0398 ± 0.0010 AU
Centre-of-mass velocity, γ	23.9 ± 0.5 km s ⁻¹	24.1 ± 0.4 km s ⁻¹
Mass ratio, $q \equiv M_2/M_1$	0.642 ± 0.043	0.625 ± 0.018
Total mass, $(M_1 + M_2) \sin^3 i$	0.0897 ± 0.0086 $M_\odot$	0.0880 ± 0.0066 $M_\odot$
Inclination, i		88.8 ± 0.1°
Primary mass, M_1		0.0541 ± 0.0046 $M_\odot$
Secondary mass, M_2		0.0340 ± 0.0027 $M_\odot$
Primary radius, R_1		0.669 ± 0.034 $R_\odot$
Secondary radius, R_2		0.511 ± 0.026 $R_\odot$
Surface temperature ratio, T_2/T_1		1.054 ± 0.006

The orbital and physical parameters of 2MASS J05352184–0546085 were determined in step-wise fashion, beginning with estimates of basic orbital parameters from the light curve. The orbital period, P , is determined directly from time-series analysis of the light curve. From geometrical considerations relating the orbital period, the relative durations of the eclipses, and their separations in time, we estimate the orbital eccentricity e and the orientation of periastron ω to be $e \approx 0.35$ and $\omega \approx 216^\circ$. With P fixed, we obtain a preliminary orbit solution by fitting the radial-velocity measurements alone, using these e and ω estimates as initial guesses. This preliminary orbit solution (column 1) yields the binary mass ratio and the total mass multiplied by $\sin^3 i$, where i is the unknown inclination of the orbital plane relative to the plane of the sky. 'Primary' and 'secondary' refer to the higher- and lower-mass brown dwarf, respectively. The preliminary orbit solution was refined, and the physical parameters of the brown dwarfs measured, by simultaneously fitting the light-curve and radial-velocity data using a standard detached-eclipsing-binary model^{8,9}. This final orbit solution (column 2) yields the orbital and physical parameters with no $\sin i$ ambiguity; i is very nearly 90° , as expected for an eclipsing geometry. From the eclipse durations and their relative depths, we also obtain the radii of the brown dwarfs and the ratio of their surface temperatures. Surprisingly, the secondary brown dwarf is warmer than the primary; this result is highly statistically significant. Uncertainties in the parameters represent standard 1σ formal errors from the covariance matrix of the fit. AU, astronomical units.

radius ($R = 0.12 R_\odot$) of the old and low-mass ($M = 0.09 M_\odot$) star in OGLE-TR-122 (ref. 22) confirms that objects with near-brown-dwarf masses do indeed have planet-like radii when they are old. Together, these observations verify the basic theoretical prediction that brown dwarfs begin their lives in a 'star-like' state—large and warm—and evolve to a more planet-like state as they contract under gravity.

At the same time, our finding that the higher-mass brown dwarf of 2MASS J05352184–0546085 is cooler than its companion is puzzling because all theoretical models predict that a brown dwarf of a given mass will at all times be warmer than a lower-mass brown dwarf of the same age. Thus, perhaps the brown dwarfs in 2MASS J05352184–0546085 are not the same age; perhaps they did not form together as a

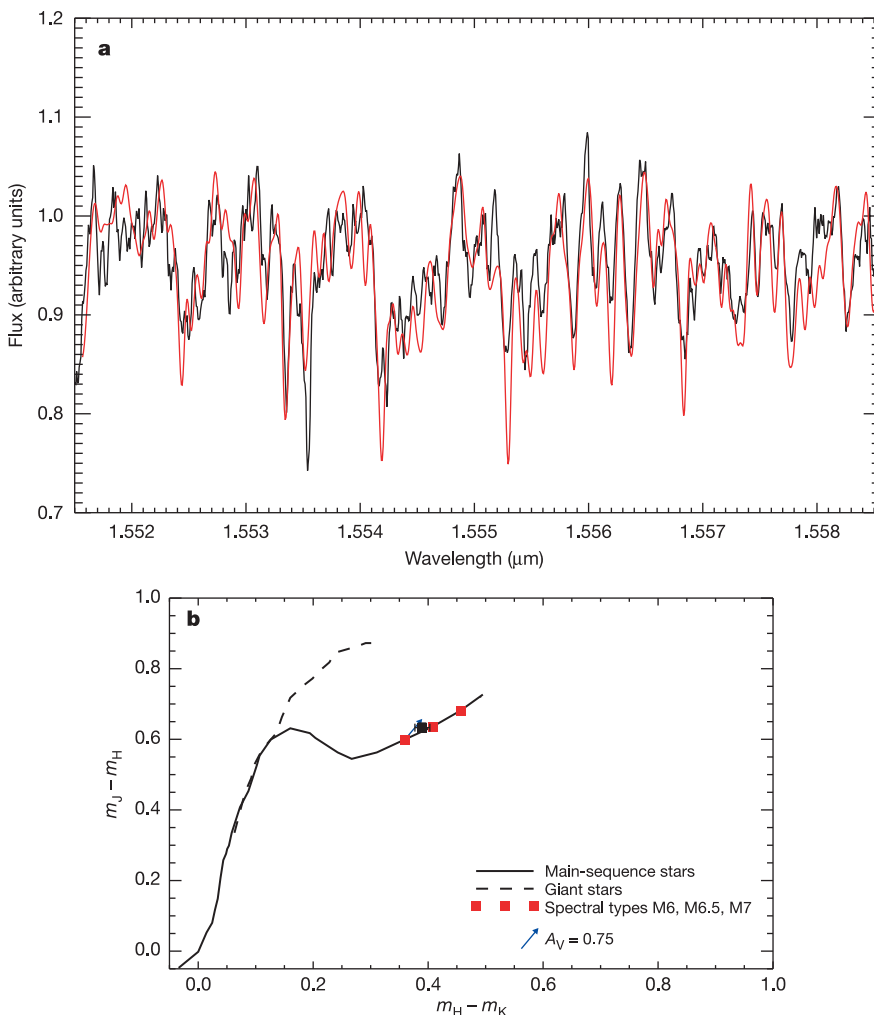


Figure 3 | Near-infrared spectrum and colours of 2MASS J05352184–0546085. **a**, The spectrum at 1.555 μm of the primary brown dwarf, obtained by averaging the eight individual spectra after shifting them to the velocity of the primary, is shown (black). For comparison, the spectrum of the best-fitting standard star—LHS 292, an old M6.5 dwarf—is also shown (red). On the basis of this very good spectral match, we adopt this standard star for the purposes of the radial-velocity analysis (Fig. 2). **b**, The near-infrared colours (J, H, Ks magnitudes, corresponding to wavelengths of 1.2, 1.6 and 2.2 μm) of 2MASS J05352184–0546085 (black symbol with error bars), are compared to standard colours of field dwarfs and giants¹⁴. The colours of 2MASS J05352184–0546085 are the mean of 30 observations from the time-series 2MASS measurements of ref. 17 and converted to the CIT photometric system as described in the 2MASS documentation. Error bars represent the uncertainty of the mean (s.e.m.) of the 30 measurements. Red symbols mark the mean colours of main-sequence spectral types M6, M6.5 and M7 (from left to right). The observed colours are consistent with no extinction, though also consistent with a spectral type of M6 seen through an extinction of $A_V = 0.75$ (blue arrow), which is not uncommon for objects in this region²⁰. We thus adopt a spectral type for the primary of $M6.5 \pm 0.5$, though systematic uncertainties may be as large as about one spectral subtype. That the colours of 2MASS J05352184–0546085 in combined light so closely match the spectral type assigned to the primary on the basis of its spectrum in **a** confirms that both brown dwarfs have very similar spectral types.

binary, but rather formed separately and were later married through a dynamical interaction. Indeed, recent theoretical work²³ and detailed numerical simulations^{24,25} suggest that dynamical interactions may be integral to the formation of brown dwarfs, but this hypothesis remains under debate²⁶. Alternatively, the theoretical model ages may simply be in error. Indeed, the model ages of low-mass stars and brown dwarfs have never been independently calibrated; instead, the models currently adopt arbitrary starting points for their calculations. 2MASS J05352184–0546085 may thus provide an empirical calibration of the initial conditions for newly formed brown dwarfs. 2MASS J05352184–0546085 may also indicate the need for additional physical ingredients in the models. For example, the presence of strong magnetic fields on one or both brown dwarfs could be affecting energy transport, thereby altering their physical structure. Interestingly, recent work has suggested that the radii of young brown dwarfs may be larger than predicted^{27–30}. Additional theoretical work will be needed to ascertain whether these ideas are plausible. In any event, the reversal of temperature with mass in 2MASS J05352184–0546085 is an unexpected result that demands explanation.

More generally, the discovery of 2MASS J05352184–0546085 represents the first direct, accurate measurement to our knowledge of the fundamental physical properties of young brown dwarfs. Detailed analyses of this unusual system promise to yield meaningful tests of the predictions of theoretical models for young brown dwarfs, and to provide rare empirical insight into the nature and origins of these ‘failed stars’.

Received 10 September 2005; accepted 2 January 2006.

- Chabrier, G. & Baraffe, I. Theory of low-mass stars and substellar objects. *Annu. Rev. Astron. Astrophys.* **38**, 337–377 (2000).
- Burrows, A., Hubbard, W. B., Lunine, J. I. & Liebert, J. The theory of brown dwarfs and extrasolar giant planets. *Rev. Mod. Phys.* **73**, 719–765 (2001).
- Basri, G. Observations of brown dwarfs. *Annu. Rev. Astron. Astrophys.* **38**, 485–519 (2000).
- Baraffe, I., Chabrier, G., Allard, F. & Hauschildt, P. H. Evolutionary models for solar metallicity low-mass stars: mass-magnitude relationships and colour-magnitude diagrams. *Astron. Astrophys.* **337**, 403–412 (1998).
- D’Antona, F. & Mazzitelli, I. Evolution of low mass stars. *Mem. Soc. Astron. It.* **68**, 807–822 (1997).
- Zapatero Osorio, M. R. et al. Dynamical masses of the binary brown dwarf GJ 569 Bab. *Astrophys. J.* **615**, 958–971 (2004).
- Andersen, J. Accurate masses and radii of normal stars. *Astron. Astrophys. Rev.* **3**, 91–126 (1991).
- Wilson, R. E. & Devinney, E. J. Realization of accurate close-binary light curves: application to MR Cygni. *Astrophys. J.* **166**, 605–620 (1971).
- Prsa, A. & Zwitter, T. A computational guide to physics of eclipsing binaries. I. Demonstrations and perspectives. *Astrophys. J.* **628**, 426–438 (2005).
- Chabrier, G., Baraffe, I., Allard, F. & Hauschildt, P. Evolutionary models for very low-mass stars and brown dwarfs with dusty atmospheres. *Astrophys. J.* **542**, 464–472 (2000).
- Scholz, A. & Eisloffel, J. Rotation and variability of very low mass stars and brown dwarfs near epsilon Ori. *Astron. Astrophys.* **429**, 1007–1023 (2005).
- Mohanty, S. & Basri, G. Rotation and activity in mid-M to L field dwarfs. *Astrophys. J.* **583**, 451–472 (2003).
- Stassun, K. G., Mathieu, R. D., Vaz, L. P. R., Stroud, N. & Vrba, F. J. Dynamical mass constraints on low-mass pre-main-sequence stellar evolutionary tracks: an eclipsing binary in Orion with a 1.0 $M_{\odot}$ primary and a 0.7 $M_{\odot}$ secondary. *Astrophys. J. Suppl.* **151**, 357–385 (2004).
- Bessell, M. S. The late-M dwarfs. *Astron. J.* **101**, 662–676 (1991).
- Slesnick, C. L., Hillenbrand, L. A. & Carpenter, J. M. The spectroscopically determined substellar mass function of the Orion Nebula cluster. *Astrophys. J.* **610**, 1045–1063 (2004).
- Hillenbrand, L. A. & White, R. J. An assessment of dynamical mass constraints on pre-main-sequence evolutionary tracks. *Astrophys. J.* **604**, 741–757 (2004).
- Carpenter, J. M., Hillenbrand, L. A. & Skrutskie, M. F. Near-infrared photometric variability of stars toward the Orion A molecular cloud. *Astron. J.* **121**, 3160–3190 (2001).
- Genzel, R. & Stutzki, J. The Orion molecular cloud and star-forming region. *Annu. Rev. Astron. Astrophys.* **27**, 41–85 (1989).
- Sicilia-Aguilar, A. et al. Accretion, kinematics, and rotation in the Orion Nebula Cluster: Initial results from Hectochelle. *Astron. J.* **129**, 363–381 (2005).
- Hillenbrand, L. A. On the stellar population and star-forming history of the Orion Nebula Cluster. *Astron. J.* **113**, 1733–1768 (1997).
- Palla, F. & Stahler, S. W. Star formation in the Orion Nebula Cluster. *Astrophys. J.* **525**, 772–783 (1999).
- Pont, F. et al. A planet-sized transiting star around OGLE-TR-122. Accurate mass and radius near the hydrogen-burning limit. *Astron. Astrophys.* **433**, L21–L24 (2005).
- Reipurth, B. & Clarke, C. The formation of brown dwarfs as ejected stellar embryos. *Astron. J.* **122**, 432–439 (2001).
- Bate, M. R. & Bonnell, I. A. The origin of the initial mass function and its dependence on the mean Jeans mass in molecular clouds. *Mon. Not. R. Astron. Soc.* **356**, 1201–1221 (2005).
- Bate, M. R., Bonnell, I. A. & Bromm, V. The formation mechanism of brown dwarfs. *Mon. Not. R. Astron. Soc.* **332**, L65–L68 (2002).
- Maxted, P. F. L. & Jeffries, R. D. On the frequency of close binary systems among very low-mass stars and brown dwarfs. *Mon. Not. R. Astron. Soc.* **362**, L45–L49 (2005).
- Mohanty, S., Jayawardhana, R. & Basri, G. Measuring fundamental parameters of substellar objects. II. Masses and radii. *Astrophys. J.* **609**, 885–905 (2004).
- Mohanty, S. et al. Measuring fundamental parameters of substellar objects. I. Surface gravities. *Astrophys. J.* **609**, 854–884 (2004).
- Doppmann, G. W. & Jaffe, D. T. A spectroscopic technique for measuring stellar properties of pre-main-sequence stars. *Astron. J.* **126**, 3030–3042 (2003).
- Doppmann, G. W., Jaffe, D. T. & White, R. J. Stellar properties of pre-main-sequence stars from high-resolution near-infrared spectra. *Astron. J.* **126**, 3043–3057 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to I. Baraffe, G. Basri, D. Golimowski, and D. Weintraub for discussions, and to D. Gudehaus and A. Prsa for software used in our analyses. This work is supported by grants to K.G.S. and R.D.M. from the National Science Foundation. The Gemini Observatory is operated by the Association of Universities for Research in Astronomy, Inc., under a cooperative agreement with the NSF on behalf of the Gemini partnership: the National Science Foundation (USA), the Particle Physics and Astronomy Research Council (UK), the National Research Council (Canada), CONICYT (Chile), the Australian Research Council (Australia), CNPq (Brazil), and CONICET (Argentina).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.G.S. (keivan.stassun@vanderbilt.edu).

Evidence for Efimov quantum states in an ultracold gas of caesium atoms

T. Kraemer¹, M. Mark¹, P. Waldburger¹, J. G. Danzl¹, C. Chin^{1,2}, B. Engeser¹, A. D. Lange¹, K. Pilch¹, A. Jaakkola¹, H.-C. Nägerl¹ & R. Grimm^{1,3}

Systems of three interacting particles are notorious for their complex physical behaviour. A landmark theoretical result in few-body quantum physics is Efimov's prediction^{1,2} of a universal set of bound trimer states appearing for three identical bosons with a resonant two-body interaction. Counterintuitively, these states even exist in the absence of a corresponding two-body bound state. Since the formulation of Efimov's problem in the context of nuclear physics 35 years ago, it has attracted great interest in many areas of physics^{3–8}. However, the observation of Efimov quantum states has remained an elusive goal^{3,5}. Here we report the observation of an Efimov resonance in an ultracold gas of caesium atoms. The resonance occurs in the range of large negative two-body scattering lengths, arising from the coupling of three free atoms to an Efimov trimer. Experimentally, we observe its signature as a giant three-body recombination loss^{9,10} when the strength of the two-body interaction is varied. We also detect a minimum^{9,11,12} in the recombination loss for positive scattering lengths, indicating destructive interference of decay pathways. Our results confirm central theoretical predictions of Efimov physics and represent a starting point with which to explore the universal properties of resonantly interacting few-body systems⁷. While Feshbach resonances^{13,14} have provided the key to control quantum-mechanical interactions on the two-body level, Efimov resonances connect ultracold matter¹⁵ to the world of few-body quantum phenomena.

Efimov's treatment of three identical bosons^{1,2} is closely linked to the concept of universality⁷ in systems with a resonant two-body interaction, where the *s*-wave scattering length *a* fully characterizes the two-body physics. When $|a|$ greatly exceeds the characteristic range ℓ of the two-body interaction potential, details of the short-range interaction become irrelevant because of the long-range nature of the wavefunction. Universality then leads to a generic behaviour in three-body physics, reflected in the energy spectrum of weakly bound Efimov trimer states. Up to now, in spite of their great fundamental importance, these states could not be observed experimentally. An observation in the realm of nuclear physics, as originally proposed by Efimov, is hampered by the presence of the Coulomb interaction, and only two-neutron halo systems with a spinless core are likely to feature Efimov states³. In molecular physics, the helium trimer is predicted to have an excited state with Efimov character⁴. The existence of this state could not be confirmed⁵. A different approach to experimentally studying the physics of Efimov states is based on the unique properties of ultracold atomic quantum gases. Such systems¹⁵ provide an unprecedented level of control, enabling the investigation of interacting quantum systems. The ultralow collision energies allow us to explore the zero-energy quantum limit. Moreover, two-body interactions can be precisely tuned on the basis of Feshbach resonances^{13,14}.

Efimov's scenario^{1,2,7} can be illustrated by the energy spectrum of the three-body system as a function of the inverse scattering length $1/a$ (Fig. 1). Let us first consider the well-known weakly bound dimer state, which only exists for large positive *a*. In the resonance regime, its binding energy is given by the universal expression $E_b = -\hbar^2/(ma^2)$, where *m* is the atomic mass and $\hbar$ is Planck's constant divided by 2π . In Fig. 1, where the resonance limit corresponds to $1/a \rightarrow 0$, the dimer energy E_b is represented by a parabola for $a > 0$. If we now add one more atom with zero energy, a natural continuum threshold for the bound three-body system (hatched line in Fig. 1) is given by the three-atom threshold ($E = 0$) for negative *a* and by the dimer-atom threshold (E_b) for positive *a*. Energy states below the continuum threshold are necessarily three-body bound states. When $1/a$ approaches the resonance from the negative-*a* side, a first Efimov trimer state appears in a range where a weakly bound two-body state does not exist. When passing through the resonance the state connects to the positive-*a* side, where it finally intersects with the dimer-atom threshold. An infinite series of such Efimov states is found when scattering lengths are increased and binding energies are decreased in powers of universal scaling

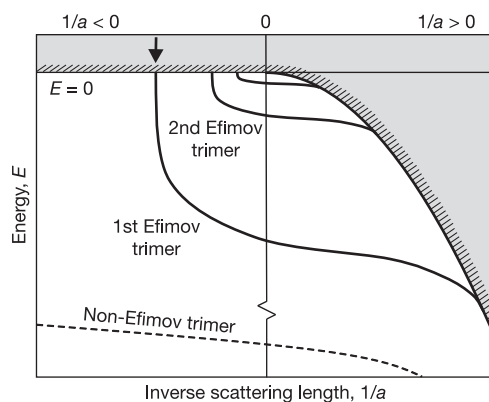


Figure 1 | Efimov's scenario. Appearance of an infinite series of weakly bound Efimov trimer states for resonant two-body interaction. The binding energy is plotted as a function of the inverse two-body scattering length $1/a$. The shaded region indicates the scattering continuum for three atoms ($a < 0$) and for an atom and a dimer ($a > 0$). The arrow marks the intersection of the first Efimov trimer with the three-atom threshold. To illustrate the series of Efimov states, we have artificially reduced the universal scaling factor from 22.7 to 2. For comparison, the dashed line indicates a tightly bound non-Efimov trimer³⁰, which does not interact with the scattering continuum.

¹Institut für Experimentalphysik, Universität Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria. ²James Franck Institute, Physics Department of the University of Chicago, 5640 S. Ellis Avenue Chicago, Illinois 60637, USA. ³Institut für Quantenoptik und Quanteninformation der Österreichischen Akademie der Wissenschaften, Otto-Hittmair-Platz 1, A-6020 Innsbruck, Austria.

factors^{1,2,7} $e^{\pi/s_0} \approx 22.7$ and $e^{-2\pi/s_0} \approx 1/515$ (where $s_0 = 1.00624$), respectively.

Resonant scattering phenomena arise as a natural consequence of Efimov's scenario¹⁶. When an Efimov state intersects with the continuum threshold at negative scattering lengths a , three free atoms in the ultracold limit resonantly couple to a trimer. This results in a triatomic Efimov resonance. At finite collision energies, the phenomenon evolves into a triatomic continuum resonance¹⁷. Another type of Efimov resonance¹⁸ is found at positive values of a for collisions between a free atom and a dimer, when Efimov states intersect with the dimer-atom threshold. While the latter type of Efimov resonance corresponds to Feshbach resonances in collisions between atoms and dimers¹⁸, triatomic Efimov resonances can be interpreted as a three-body generalization to Feshbach resonances⁸.

Striking manifestations of Efimov physics have been predicted for three-body recombination processes in ultracold gases with tunable two-body interactions^{7,9–12,19}. Three-body recombination leads to losses from a trapped gas with a rate proportional to the third power of the atomic number density. These losses are commonly described²⁰ in terms of a loss rate coefficient L_3 . In the resonant case ($|a| \gg \ell$), it is convenient to express this coefficient in the form $L_3 = 3C(a)\hbar a^4/m$, separating a general a^4 -scaling^{20,21} from an additional dependence^{9,10,12} $C(a)$. Efimov physics is reflected in a logarithmically periodic behaviour $C(22.7a) = C(a)$, corresponding to the scaling of the infinite series of weakly bound trimer states. For negative scattering lengths, the resonant coupling of three atoms to an Efimov state opens up fast decay channels into deeply bound dimer states plus a free atom.

Triatomic Efimov resonances thus show up in giant recombination loss. This striking phenomenon was first identified in numerical solutions to the adiabatic hyperspherical approximation of the three-body Schrödinger equation, assuming simple model potentials and interpreted in terms of tunnelling through a potential barrier in the three-body entrance channel⁹. A different theoretical approach^{7,10}, based on effective field theory, provides the analytic expression $C(a) = 4,590 \sinh(2\eta_-) / (\sin^2[s_0 \ln(|a|/a_-)] + \sinh^2 \eta_-)$. The free parameter a_- for the resonance positions at a_- , $22.7 a_-$, ... depends on the short-range part of the effective three-body interaction and is thus not determined in the frame of the universal long-range theory. As a second free parameter, the dimensionless quantity η_- describes the unknown decay rate of Efimov states into deeply bound dimer states plus a free atom, and thus characterizes the resonance width.

Our measurements are based on the magnetically tunable interaction properties of caesium atoms²² in the lowest internal state. By applying fields between 0 and 150 G, we varied the s -wave scattering length a in a range between $-2,500a_0$ to $1,600a_0$, where a_0 is Bohr's radius. Accurate three-body loss measurements are facilitated by the fact that inelastic two-body loss is energetically forbidden²⁰. The characteristic range of the two-body potential is given by the van der Waals length²³, which for caesium is $\ell \approx 100a_0$. This leaves us with enough room to study the universal regime requiring $|a| \gg \ell$. For negative a , a maximum value of 25 is attainable for $|a|/\ell$. Efimov's estimate $\frac{1}{\pi} \ln(|a|/\ell)$ for the number of weakly bound trimer states² suggests the presence of one Efimov resonance in the accessible range of negative scattering lengths.

Our experimental results (Fig. 2), obtained with optically trapped thermal samples of caesium atoms in two different set-ups (see Methods), indeed show a giant loss feature marking the expected resonance. We present our data in terms of a recombination length⁹ $\rho_3 = [2m/(\sqrt{3}\hbar)L_3]^{1/4}$, which leads to the simple relation $\rho_3/a = 1.36C^{1/4}$. Note that the general a^4 -scaling corresponds to a linear behaviour in $\rho_3(a)$ (straight lines in Fig. 2). A fit of the analytic theory^{7,10} to our experimental data taken for negative a at temperatures $T \approx 10$ nK shows a remarkable agreement and determines the resonance position to $a_- = -850(20)a_0$ and the decay parameter to $\eta_- = 0.06(1)$. The pronounced resonance behaviour with a small value for the decay parameter ($\eta_- \ll 1$) demonstrates a sufficiently

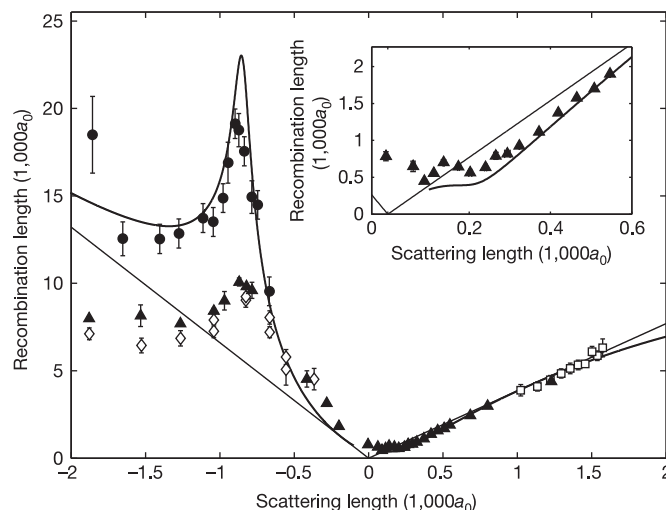


Figure 2 | Observation of the Efimov resonance in measurements of three-body recombination. The recombination length $\rho_3 \propto L_3^{1/4}$ is plotted as a function of the scattering length a . The dots and the filled triangles show the experimental data from set-up A for initial temperatures around 10 nK and 200 nK, respectively. The open diamonds are from set-up B at temperatures of 250 nK. The open squares are previous data²⁰ at initial temperatures between 250 and 450 nK. The solid curve represents the analytic model from effective field theory⁷ with $a_- = -850a_0$, $a_+ = 1,060a_0$, and $\eta_- = \eta_+ = 0.06$. The straight lines result from setting the $\sin^2$ and $\cos^2$ -terms in the analytic theory to 1, which gives a lower recombination limit for $a < 0$ and an upper limit for $a > 0$. The inset shows an expanded view for small positive scattering lengths with a minimum for $C(a) \propto (\rho_3/a)^4$ near $210a_0$. The displayed error bars refer to statistical uncertainties only. Uncertainties in the determination of the atomic number densities may lead to additional calibration errors for ρ_3 of up to 20%.

long lifetime of Efimov trimers to allow their observation as distinct quantum states.

All the results discussed so far are valid in the zero-energy collision limit of sufficiently low temperatures. For ultralow but non-zero temperatures the recombination length is unitarity limited¹⁹ to $5.2\hbar/(mk_B T)^{1/2}$. For $T = 10$ nK this limit corresponds to about $60,000a_0$ and our sample is thus cold enough to justify the zero-temperature limit. For 250 nK, however, unitarity limits the recombination length to about $12,000a_0$. The Efimov resonance is still visible at temperatures of 200 and 250 nK (filled triangles and open diamonds in Fig. 2). The slight shift to lower values of $|a|$ suggests the evolution of the zero-energy Efimov resonance into a triatomic continuum resonance¹⁷. In further experiments at higher temperatures (data not shown) we observed the resonance to disappear above ~ 500 nK.

For positive scattering lengths, we found three-body losses to be typically much weaker than for negative values. Our measurements are consistent with a maximum recombination loss of $C(a) \approx 70$, or equivalently $\rho_3 \approx 3.9a$, as predicted by different theories^{9,11,12} (straight line for $a > 0$ in Fig. 2). For a below $600a_0$ the measured recombination length significantly drops below this upper limit (inset in Fig. 2). The analytic expression from effective field theory^{7,12} for $a > 0$ reads $C(a) = 67.1e^{-2\eta_+} (\cos^2[s_0 \ln(a/a_+)] + \sinh^2 \eta_+) + 16.8(1 - e^{-4\eta_+})$ with the two free parameters a_+ and η_+ . The first term describes recombination into the weakly bound dimer state with an oscillatory behaviour that is due to an interference effect between two different pathways^{9,11}. The second term results from decay into deeply bound states. We use this expression to fit our data points with $a > 5\ell \approx 500a_0$. This somewhat arbitrary condition is introduced as a reasonable choice to satisfy $a \gg \ell$ for the validity of the universal theory. The fit is quite insensitive to the value of the decay parameter and yields $\eta_+ < 0.2$. This result is consistent with the theoretical assumption¹⁰ of the same value for the decay

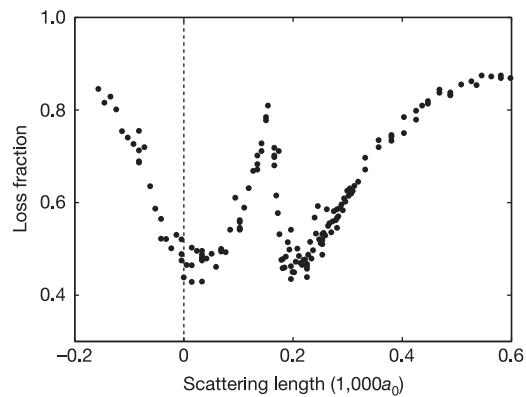


Figure 3 | Atom loss for small scattering lengths. Besides a minimum near zero scattering length, we identify a minimum of recombination loss at $\sim 210a_0$, which can be attributed to a predicted destructive interference effect^{9,11,12}.

parameter for positive and negative a , which in our case is $\eta_+ = \eta_- = 0.06$. For maximum $C(a)$, we obtain $a_+ = 1,060(70)a_0$. According to theory⁷, the trimer state hits the dimer-atom threshold at $a = 1.1a_+ \approx 1,170a_0$. The logarithmic periodicity of the Efimov scenario suggests that adjacent loss minima occur at $\sqrt{22.7} \times 1,060a_0 \approx 5,000a_0$ and at $1,060a_0/\sqrt{22.7} \approx 220a_0$. While the former value is out of our accessible range, the latter value ($a \approx 2\ell$) is too small to strictly justify universal behaviour in the resonance limit ($a \gg \ell$). Nevertheless, our experimental results (inset to Fig. 2) indicate a minimum at $a \approx 210a_0$ and the analytic expression for $C(a)$ is found to describe our data quite well down to this minimum.

The occurrence of the interference minimum in three-body loss is demonstrated more clearly in another set of experiments (Fig. 3), where we simply measured the loss of atoms after a fixed storage time in the optical trap. This minimum is located at $a = 210(10)a_0$ in addition to a second minimum close to zero scattering length. We point out that the existence of the minimum at $210a_0$ is very advantageous for efficient evaporative cooling of caesium as it combines a large scattering cross-section with very low loss. Inadvertently, we have already benefited from this loss minimum for the optimized production of a Bose-Einstein condensate of caesium²⁴.

The comparison of our experimental results to available three-body theory shows remarkable agreement, although the collision physics of caesium is in general a very complicated multi-channel scattering problem. We believe that the particular nature of the broad, “open-channel dominated” Feshbach resonance²⁵ that underlies the tunability of our system plays a crucial role. For such a resonance, the two-body scattering problem can be described in terms of an effective single-channel model. It is very interesting to investigate to what degree this great simplification of the two-body physics extends to the three-body problem. Here we particularly wonder how the regions of positive and negative scattering lengths are connected in our experiment, where a is changed through a zero crossing—that is, through a non-universal region, and not across the universal resonance region.

In our case, there is no obvious connection between the Efimov state that leads to the observed resonance for $a < 0$ and the states responsible for the behaviour for $a > 0$. In our analysis of the experimental data, we have thus independently fitted the data sets for negative and positive a . Nevertheless, the resulting values for the two independent fit parameters a_- and a_+ do suggest a connection: for the ratio $a_+/|a_-|$ our experiment yields 1.25(9), whereas universal theory⁷ predicts 0.96(3). These numbers are quite close in view of the Efimov factor of 22.7. If it is not an accidental coincidence, we speculate that the apparent relation between a_+ and a_- may be a further consequence of universality in a system where the resonant two-body interaction can be modelled in terms of a single scattering

channel. In general, the multi-channel nature of three-body collisions near Feshbach resonances^{26,27} leads to further interesting questions, such as whether there may be resonance effects beyond the Efimov scenario. Advances in three-body theory are necessary to answer these questions and to provide a complete interpretation of our present observations.

In the past few years, applications of Feshbach resonances in ultracold gases and the resulting ability to create dimer states have set the stage for many new developments in matter-wave quantum physics. The observation of an Efimov resonance now confirms the existence of weakly bound trimer states and opens up new ways^{6,8} of experimentally exploring the intriguing physics of few-body quantum systems.

METHODS

Magnetic tuning of the two-body interaction. For Cs atoms in their energetically lowest state (quantum numbers $F = 3$ for the total spin and $m_F = 3$ for its projection) the s -wave scattering length a varies strongly with the magnetic field²². Between 0 and 150 G the dependence can in general be well approximated by the fitting formula:

$$a(B)/a_0 = (1,722 + 1.52B/G) \left(1 - \frac{28.72}{B/G + 11.74} \right)$$

except for a few narrow Feshbach resonances²². The smooth variation of the scattering length in the low-field region results from a broad Feshbach resonance centred at about -12 G (equivalent to $+12$ G in the state $F = 3, m_F = -3$). In all our measurements we excluded the magnetic field regions where the narrow Feshbach resonances influence the scattering behaviour through coupling to other molecular potentials. The Efimov resonance is centred at 7.5 G.

Trap set-ups and preparation of the Cs gases. All measurements were performed with trapped thermal samples of caesium atoms at temperatures T ranging from 10 to 250 nK. We used two different experimental set-ups, which have been described elsewhere^{24,28}.

In set-up A we first produced an essentially pure Bose-Einstein condensate with up to 250,000 atoms in a far-detuned crossed optical dipole trap generated by two 1,060-nm Yb-doped fibre laser beams²⁴. We then ramped the magnetic field to 16.2 G, where the scattering length is negative with a value of $-50a_0$, thus inducing a collapse of the condensate²⁹. After an equilibration time of 1 s we were left with a thermal sample at typically $T = 10$ nK containing up to 20,000 atoms at peak densities ranging from $n_0 = 3 \times 10^{11} \text{ cm}^{-3}$ to $3 \times 10^{12} \text{ cm}^{-3}$. Alternatively, we interrupted the evaporation process before condensation to produce thermal samples at $T \approx 200$ nK in a crossed dipole trap generated by one of the 1,060-nm beams and a 10.6- μm CO₂ laser beam. After recompression of the trap this produced typical densities of $n_0 = 5 \times 10^{13} \text{ cm}^{-3}$. The measurements in the region of the loss minima as displayed in Fig. 3 were taken after a storage time of 200 ms at initial densities of $n_0 = 6 \times 10^{13} \text{ cm}^{-3}$.

In set-up B we used an optical surface trap²⁸ in which we prepared a thermal sample of 10,000 atoms at $T \approx 250$ nK via forced evaporation at a density of $n_0 = 1.0 \times 10^{12} \text{ cm}^{-3}$. The dipole trap was formed by a repulsive evanescent laser wave on top of a horizontal glass prism in combination with a single horizontally confining 1,060-nm laser beam propagating along the vertical direction.

Determination of three-body loss rate coefficients. We measured three-body loss rates in set-up A by recording the time evolution of the atom number N and the temperature T . A detailed description of this procedure has been given in ref. 20. In brief, the process of three-body recombination not only leads to a loss of atoms, but also induces ‘anti-evaporation’ and recombination heating. The first effect is present at any value of the scattering length a . The second effect occurs for positive values of a when the recombination products remain trapped. Atom loss and temperature rise are modelled by a set of two coupled nonlinear differential equations. We used numerical solutions to this set of equations to fit our experimental data. From these fits, together with measurements of the trapping parameters, we obtained the rate coefficient L_3 . In set-up B we recorded the loss at decay times sufficiently short to make sure that heating is negligible.

Received 12 December 2005; accepted 2 February 2006.

1. Efimov, V. Energy levels arising from resonant two-body forces in a three-body system. *Phys. Lett. B.* **33**, 563–564 (1970).
2. Efimov, V. Weakly-bound states of three resonantly-interacting particles. *Sov. J. Nucl. Phys.* **12**, 589–595 (1971).
3. Jensen, A. S., Riisager, K., Fedorov, D. V. & Garrido, E. Structure and reactions of quantum halos. *Rev. Mod. Phys.* **76**, 215–261 (2004).

4. Lim, T. K., Duffy, K. & Damer, W. C. Efimov state in the ^4He trimer. *Phys. Rev. Lett.* **38**, 341–343 (1977).
5. Brühl, R. *et al.* Matter wave diffraction from an inclined transmission grating: Searching for the elusive ^4He trimer Efimov state. *Phys. Rev. Lett.* **95**, 063002 (2005).
6. Braaten, E., Hammer, H.-W. & Kusunoki, M. Efimov states in a Bose-Einstein condensate near a Feshbach resonance. *Phys. Rev. Lett.* **90**, 170402 (2003).
7. Braaten, E. & Hammer, H.-W. Universality in few-body systems with large scattering length. Preprint at (<http://arXiv.org/abs/cond-mat/0410417>) (2004).
8. Stoll, M. & Köhler, T. Production of three-body Efimov molecules in an optical lattice. *Phys. Rev. A* **72**, 022714 (2005).
9. Esry, B. D., Greene, C. H. & Burke, J. P. Jr Recombination of three atoms in the ultracold limit. *Phys. Rev. Lett.* **83**, 1751–1754 (1999).
10. Braaten, E. & Hammer, H.-W. Three-body recombination into deep bound states in a Bose gas with large scattering length. *Phys. Rev. Lett.* **87**, 160407 (2001).
11. Nielsen, E. & Macek, J. H. Low-energy recombination of identical bosons by three-body collisions. *Phys. Rev. Lett.* **83**, 1566–1569 (1999).
12. Bedaque, P. F., Braaten, E. & Hammer, H.-W. Three-body recombination in Bose gases with large scattering length. *Phys. Rev. Lett.* **85**, 908–911 (2000).
13. Tiesinga, E., Verhaar, B. J. & Stoof, H. T. C. Threshold and resonance phenomena in ultracold ground-state collisions. *Phys. Rev. A* **47**, 4114–4122 (1993).
14. Inouye, S. *et al.* Observation of Feshbach resonances in a Bose-Einstein condensate. *Nature* **392**, 151–154 (1998).
15. Southwell, K. (ed.) *Ultracold matter Nature (Insight)* **416**, 205–246 (2002).
16. Efimov, V. Low-energy properties of three resonantly interacting particles. *Sov. J. Nucl. Phys.* **29**, 546–553 (1979).
17. Bringas, F., Yamashita, M. T. & Frederico, T. Triatomic continuum resonances for large negative scattering lengths. *Phys. Rev. A* **69**, 040702(R) (2004).
18. Nielsen, E., Suno, H. & Esry, B. D. Efimov resonances in atom-diatom scattering. *Phys. Rev. A* **66**, 012705 (2002).
19. D'Incao, J. P., Suno, H. & Esry, B. D. Limits on universality in ultracold three-boson recombination. *Phys. Rev. Lett.* **93**, 123201 (2004).
20. Weber, T., Herbig, J., Mark, M., Nägerl, H.-C. & Grimm, R. Three-body recombination at large scattering lengths in an ultracold atomic gas. *Phys. Rev. Lett.* **91**, 123201 (2003).
21. Fedichev, P. O., Reynolds, M. W. & Shlyapnikov, G. V. Three-body recombination of ultracold atoms to a weakly bound s level. *Phys. Rev. Lett.* **77**, 2921–2924 (1996).
22. Chin, C. *et al.* Precision Feshbach spectroscopy of ultracold Cs_2 . *Phys. Rev. A* **70**, 032701 (2004).
23. Bolda, E. L., Tiesinga, E. & Julienne, P. S. Effective-scattering-length model of ultracold atomic collisions and Feshbach resonances in tight harmonic traps. *Phys. Rev. A* **66**, 013403 (2002).
24. Kraemer, T. *et al.* Optimized production of a cesium Bose-Einstein condensate. *Appl. Phys. B* **79**, 1013–1019 (2004).
25. Köhler, T., Goral, K. & Julienne, P. S. Production of cold molecules via magnetically tunable Feshbach resonances. Preprint at (<http://arxiv.org/abs/cond-mat/0601420>) (2006).
26. Kartavtsev, O. I. & Macek, J. H. Low-energy three-body recombination near a Feshbach resonance. *Few-Body Syst.* **31**, 249–254 (2002).
27. Petrov, D. S. Three-boson problem near a narrow Feshbach resonance. *Phys. Rev. Lett.* **93**, 143201 (2004).
28. Rychtarik, D., Engeser, B., Nägerl, H.-C. & Grimm, R. Two-dimensional Bose-Einstein condensate in an optical surface trap. *Phys. Rev. Lett.* **92**, 173003 (2004).
29. Donley, E. A. *et al.* Dynamics of collapsing and exploding Bose-Einstein condensates. *Nature* **412**, 295–299 (2001).
30. Thomas, L. H. The interaction between a neutron and a proton and the structure of H^3 . *Phys. Rev.* **47**, 903–909 (1935).

Acknowledgements We thank E. Braaten, C. Greene, B. Esry, H. Hammer and T. Köhler for many discussions and E. Kneringer for support regarding the data analysis. We acknowledge support by the Austrian Science Fund (FWF) within Spezialforschungsbereich 15 and within the Lise Meitner programme, and by the European Union in the frame of the TMR networks 'Cold Molecules' and 'FASTNet'. M.M. is supported within the Doktorandenprogramm of the Austrian Academy of Sciences.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.-C.N. (christoph.naegerl@ultracold.at).

Visualizing dislocation nucleation by indenting colloidal crystals

Peter Schall^{1,†}, Itai Cohen^{1,2,†}, David A. Weitz^{1,2} & Frans Spaepen¹

The formation of dislocations is central to our understanding of yield, work hardening, fracture, and fatigue¹ of crystalline materials. While dislocations have been studied extensively in conventional materials, recent results have shown that colloidal crystals offer a potential model system for visualizing their structure and dynamics directly in real space². Although thermal fluctuations are thought to play a critical role in the nucleation of these defects, it is difficult to observe them directly. Nano-indentation, during which a small tip deforms a crystalline film, is a common tool for introducing dislocations into a small volume that is initially defect-free^{3–10}. Here, we show that an analogue of nano-indentation performed on a colloidal crystal provides direct images of defect formation in real time and on the single particle level, allowing us to probe the effects of thermal fluctuations. We implement a new method to determine the strain tensor of a distorted crystal lattice and we measure the critical dislocation loop size and the rate of dislocation nucleation directly. Using continuum models, we elucidate the relation between thermal fluctuations and the applied strain that governs defect nucleation. Moreover, we estimate that although bond energies between particles are about fifty times larger in atomic systems, the difference in attempt frequencies makes the effects of thermal fluctuations remarkably similar, so that our results are also relevant for atomic crystals.

To reduce strain energies, colloidal crystals nucleate dislocations, which mark the boundary of a surface along which the crystal has been uniformly sheared¹. Dislocation nucleation reflects a competition between the energy cost for creating a dislocation loop and the energy gain from the strain relieved by the shear along the surface bounded by the loop. The nucleation energy reaches a maximum, U_c , at the critical loop radius, r_c . Therefore, dislocations are formed only if the loops become greater than r_c . The energy barrier for defect nucleation depends sensitively on the applied strain; at low strains, thermal fluctuations are the only means of overcoming this barrier. While recent *in situ* techniques allow the imaging of the evolution of dislocations on a medium-range length scale during indentation using transmission electron microscopy^{9,10}, the nucleation of dislocation loops and the consequences of thermal effects are extremely difficult to observe directly in atomic systems.

We grow a 43- μm -thick face-centred cubic (f.c.c.) crystal in the [100] direction by slowly sedimenting 1.55- μm -diameter silica particles onto a patterned [100] substrate¹¹. The silica particles are suspended in a mixture of water and dimethylsulphoxide (DMSO), which matches their refractive index. We add a small amount of fluorescein to the solvent so that under fluorescence the particles appear as dark spots on a bright background. By varying the crystal thickness we confirm that in this crystal sufficient thermally induced particle motion persists that the effects of thermal fluctuations can

still be observed on reasonable timescales. We indent the crystal with a sewing needle (Singer sewing needles, 25 assorted sharps) attached to a piezoelectric translation stage. The needle has an almost-hemispherical tip with a diameter of 40 μm ; this tip shape induces a strain field whose maximum shear strain is well below the contact surface, and is commonly used for studying dislocation nucleation. The length scale ratios between tip diameter, particle radius, and film thickness are similar to the ratios used in nano-indentation experiments; thus the experiments are comparable.

We use laser diffraction microscopy (LDM)² to obtain an overview of the dislocation nucleation process. The incident beam forms a 36°

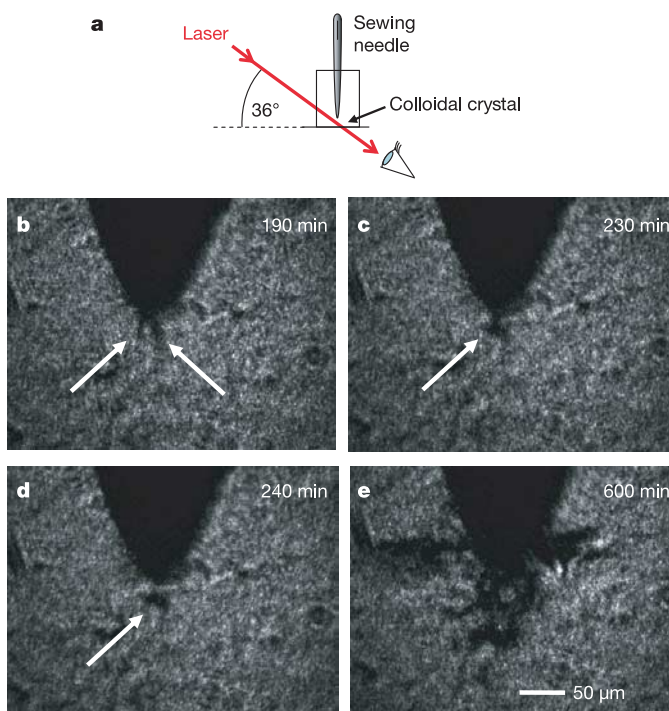


Figure 1 | Laser diffraction microscopy images of defect structure. **a**, Schematic of the indentation configuration. The imaging line of sight (red line) forms a 36° angle with the template and corresponds to the [111] direction of the f.c.c. lattice. **b–e**, LDM images depicting dark regions that correspond to crystal lattice distortions. **b**, Arrows indicate two dark regions observed at $t = 190$ min. These regions first appear at $t = 160$ min and persist until $t = 220$ min. **c, d**, Arrows indicate a dislocation loop that becomes detached from the needle. **e**, Final dislocation structure at $t = 600$ min.

¹Division of Engineering and Applied Sciences, ²Department of Physics, Harvard University, 9 Oxford Street, Cambridge, Massachusetts 02138, USA. [†]Present addresses: Van der Waals-Zeeman Institute, University of Amsterdam, Valckenierstraat 65, 1018 XE Amsterdam, The Netherlands (P.S.); Department of Physics, Cornell University, 508 Clark Hall, Ithaca, NY 14853, USA (I.C.).

angle with the plane of the template and is parallel to the $[111]$ direction of the f.c.c. crystal (Fig. 1a). We use the transmitted beam for imaging the defect structure. LDM exploits the difference in scattering between the perfect and the distorted crystal lattices near the dislocation core, and produces a real-space image in which dislocations appear as dark lines on a bright background. We lower the needle at a rate of $3.4 \mu\text{m h}^{-1}$ and record LDM images (see Supplementary Information video). Surprisingly, even before the needle touches the crystal, thermal effects produce local intensity fluctuations which last several seconds.

Four images of the indented crystal are shown in Fig. 1. About 190 min after we start lowering the needle, we observe two dark regions beneath the needle tip (arrows in Fig. 1b), which we associate with the strained lattice. Remarkably, these regions exhibit intensity fluctuations that last several minutes. This suggests that, in the strained lattice, thermal fluctuations become more spatially correlated, and persist over longer timescales than in the unstrained lattice. After 230 min, we observe a dark circular spot, about $8 \mu\text{m}$ in radius, centred $15 \mu\text{m}$ below the needle (Fig. 1c). This spot grows in size and finally detaches from the needle (Fig. 1d). These images show the nucleation of a dislocation loop and its eventual detachment from the needle. After the first event, we observe nucleation and propagation of many more loops, which form a complex dislocation network after 600 min (Fig. 1e). Because of the complexity of this network, its three-dimensional evolution needs to be explored on the particle scale.

We use confocal microscopy to image individual particles in a $60 \mu\text{m}$ by $60 \mu\text{m}$ by $30 \mu\text{m}$ section of the colloidal crystal (Fig. 2a). This technique allows us to image the dislocations and to determine the strain field caused by the indentation. We determine individual particle positions in three dimensions with an accuracy of about $0.03 \mu\text{m}$ in the x and y directions and $0.05 \mu\text{m}$ in the z direction¹². In

this experiment, we start with the needle tip about $2 \mu\text{m}$ above the crystal surface and lower it at a rate of $2.9 \mu\text{m h}^{-1}$. After 98 min, we observe a pronounced fluctuation in the crystal lattice below the needle tip, where particle rows spread apart, trying to incorporate a defect. A defect is clearly visible, as indicated by the blue arrow in the confocal microscope image of the particle configuration at $z = 15 \mu\text{m}$ below the tip, shown in Fig. 2b. The extent of the defect is indicated by the black arrows. Remarkably, the particle rows close again after about 5 min. After five such fluctuations appear and disappear, a stable defect finally nucleates at 112 min (Fig. 2c) and grows (Fig. 2d).

To determine the full structure of the defects, we show three-dimensional reconstructions of the unstable and stable defects in Figs 2e–g. The x , y and z axes correspond to the $(1\bar{1}0)$, (110) , and (001) directions of the f.c.c. lattice, respectively. At the defects, three of the six opposing nearest neighbours exhibit distorted ‘bond’ angles that are smaller than 180° . To visualize this, we use yellow spheres to indicate opposing particles at angles between 165° and 155° , and red spheres to indicate opposing particles at angles less than 155° . Red particles accumulate roughly $15 \mu\text{m}$ below the needle tip. They mark a surface along the hexagonal close-packed (h.c.p.) plane where the crystal has changed its stacking order, confirming that this is a planar defect. The yellow particles that lie at the boundary of this stacking fault in Fig. 2g trace the dislocation line. The Burgers vector of the dislocation is oriented along the $(1\bar{1}\bar{2})$ direction and has a length of $0.94 \mu\text{m}$, which corresponds to the distance between adjacent wells of the h.c.p. plane. This defines the dislocation glide plane, on which defects nucleate most easily owing to the shallow potential wells. This type of dislocation is known as a Shockley partial dislocation and is the most prominent dislocation observed in f.c.c. metals. Unlike metals, where Shockley partial dislocations typically appear in pairs to reduce the extent of the stacking fault, in our

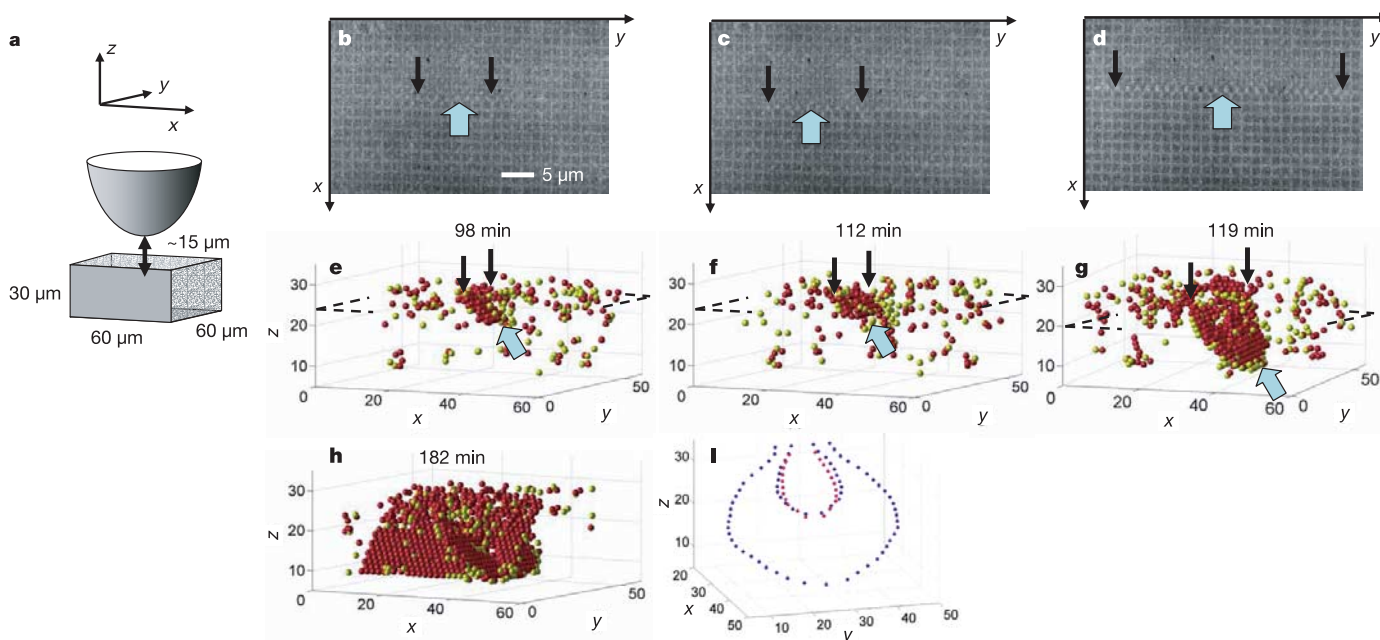


Figure 2 | Defect formation on the particle scale. **a**, Schematic showing the position of the needle tip with respect to the $60 \mu\text{m}$ by $60 \mu\text{m}$ by $30 \mu\text{m}$ crystal section depicted in **b** to **i**. **b–d**, Confocal microscope images of defect nucleation, taken at about $15 \mu\text{m}$ below the needle tip. **b**, Arrows indicate a defect that disappears after several minutes. **c**, Arrows indicate a defect that is stable and grows in size. **d**, Same defect as in **c** after it has fully grown. **e–g**, Reconstructed images of defect formation correspond to the confocal microscope images **b–d**. Yellow and red spheres indicate particles with slightly and highly distorted nearest-neighbour configurations, respectively. Arrows indicate stacking faults that are bound by a Shockley partial

dislocation. Dashed lines at $z \approx 20 \mu\text{m}$ indicate the height at which the confocal images are taken. **h**, Reconstruction shows a second stable defect, which nucleates on an adjacent intersecting plane. This defect appears at $t = 154$ min. **i**, Traces of the dislocation lines that surround the defects depicted in **e–g**. The red dots trace the dislocation line that surrounds the defect in **e**. The blue dots trace the dislocation lines that surround the stable defect shown in **f** and **g**. These traces were determined from the raw confocal images such as those in **b–d** by pinpointing the intersection of the dislocation line (black arrows in **b–d**) with the image planes.

colloidal crystal these dislocations appear to be unpaired. This difference results from the vanishingly low energy cost associated with stacking faults in hard-sphere crystals. A second dislocation is nucleated on an adjacent intersecting plane at $t = 182$ min, as shown in Fig. 2h.

These microscopic observations provide a unique opportunity for direct measurement of the critical radius, r_c . We trace the dislocation lines that surround the defects shown in Figs 2e–g by carefully examining the confocal images (Fig. 2i legend). The traces reveal circular loop shapes that open towards the crystal surface (Fig. 2i). The radius of the unstable defect (Fig. 2i, red dots) is approximately $5\ \mu\text{m}$ while the radius of the stable defect (Fig. 2i, blue dots, small loop) is $7\ \mu\text{m}$; thus, $r_c \approx 6\ \mu\text{m}$.

To explore the mechanism that drives the nucleation of these defects, we determine the components of the applied strain along the h.c.p. planes, where the defects nucleate. The elastic strain tensor is determined from the distortion in the configurations of the nearest neighbours of the particles. For each particle, we compare the vectors $\mathbf{d}_i$ to its 12 nearest neighbours i with the nearest-neighbour vectors $\mathbf{D}_i$ of the ideal f.c.c. lattice. The best affine deformation tensor, α , that transforms the ideal vectors, $\mathbf{D}_i$, to the real vectors, $\mathbf{d}_i$, is determined by minimizing the mean square difference $\sum_i (\mathbf{d}_i - \alpha \mathbf{D}_i)^2$ (ref. 13). The symmetric part of α corresponds to the local strain tensor of the particle under consideration. To smooth the resulting strain distribution, we average the strain tensor for each particle with those of its nearest neighbours and assign the resultant average strain tensor, ε_{ij} , to the particle under consideration.

The advantage of using confocal microscopy to image individual particles and measure the full strain tensor is that we can determine γ , the shear component of the strain on the $(1\ \bar{1}\ 1)$ plane where the first defect nucleates. To implement this new method of determining the shear strain, we calculate the average strain tensor ε'_{ij} in a rotated coordinate system, in which the x' and z' axes align with the $(1\ \bar{1}\ 2)$ and $(1\ \bar{1}\ 1)$ directions, respectively. We take $\gamma = 2\varepsilon'_{x'z'}$. We illustrate

various cuts through the crystal, and represent the magnitude of γ with colour. The γ distribution shortly before the defect in Fig. 2e nucleates is shown in Fig. 3a and b. A region of high positive shear strain (red) is centred approximately $10\ \mu\text{m}$ below the needle, where the average value of γ is approximately 0.06. Strikingly, the shape of this strain distribution matches the contrast profile observed with LDM (Fig. 1b). Defect nucleation occurs entirely in this region of high shear strain.

The strain field changes drastically after the first dislocation nucleates. To highlight this change, in Fig. 3c we plot the γ distribution along the y – z plane that intersects the dislocation loop along its diameter and is perpendicular to the cut shown in Fig. 3a. The negative shear strain induced by the dislocation loop (blue) reduces the cumulative strain that drives nucleation of a second dislocation on the same glide plane. In contrast, the distribution of γ' , the component of the strain on the (111) plane, still has a region of high positive strain (Fig. 3d). This results in the nucleation of a second dislocation loop in an adjacent, but intersecting h.c.p. plane (Fig. 2h).

The distribution of γ in Fig. 3c shows that the negative shear strain regions (blue) inside the dislocation loop almost overlap. Because the contrast in the LDM images arises from these lattice distortions, this observation validates our interpretation that small dislocation loops appear as dark spots in Fig. 1c and d. This strain distribution also allows us to determine the precise position of the dislocation loop and where it intersects the y – z plane in Fig. 3c. To accomplish this, we focus on particles within $1.6\ \mu\text{m}$ of the dashed line and plot γ versus y in Fig. 3e. The strain diverges on both sides of the dislocation with the divergence truncated at the dislocation core. The crossovers at $y = 7$ and $y = 45$, where γ switches sign, demarcate precisely the intersection of the dislocation with the y – z plane. This is in excellent agreement with the location of the dislocation line shown in Fig. 2i.

The precision of our measurements presents an opportunity to make direct comparisons with continuum models^{14–16}. We consider

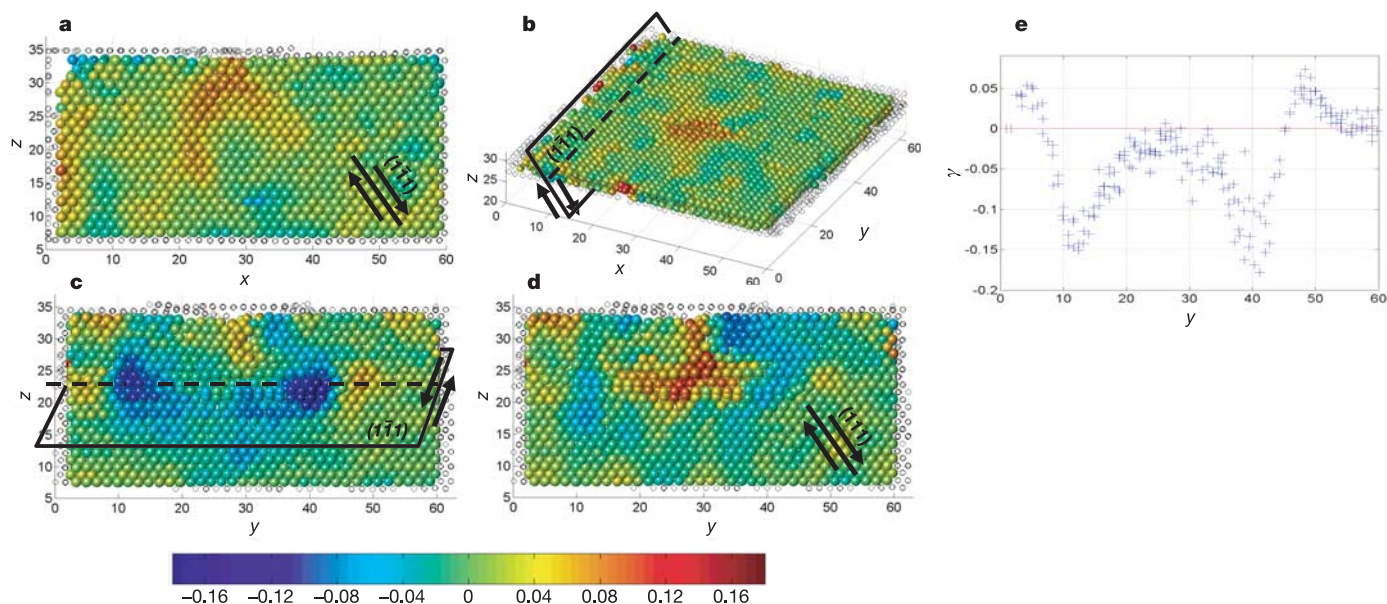


Figure 3 | Strain distribution in the indented crystalline film. Particle colour in **a–d** indicates the value of the local shear strain (see colour scale). Positive shear strain is indicated by arrows in each figure. **a, b**, Two cuts showing the distribution of γ , the shear strain component on the $(1\ \bar{1}\ 1)$ plane along $[1\ \bar{1}\ 2]$ at $t = 91$ min. **a**, $3\text{-}\mu\text{m}$ -thick x – z crystal section centred below the needle at $y = 29\ \mu\text{m}$. **b**, $3\text{-}\mu\text{m}$ -thick x – y crystal section located $10\ \mu\text{m}$ below the needle at $z = 29\ \mu\text{m}$. **c, d**, Two cuts showing the strain distribution at $t = 119$ min, shortly after the first dislocation loop nucleated. Both figures show a $3\text{-}\mu\text{m}$ -thick y – z section at $x = 33\ \mu\text{m}$, which intersects

the loop at a diametrical plane. **c**, Distribution of γ , the component of the shear strain on the defect plane. Adjacent regions of high negative shear strain (blue particles) and moderately positive shear strain (yellow particles) are observed. **d**, Distribution of γ' , the shear strain component on the intersecting $\{111\}$ plane. In contrast to **c**, a central region of high positive shear strain is observed below the needle tip. **e**, Shear strain values versus y for particles within $1.6\ \mu\text{m}$ of the dashed line at $z = 23\ \mu\text{m}$ in **c**. γ switches sign at $y = 7\ \mu\text{m}$ and $y = 45\ \mu\text{m}$. This demarcates where the dislocation line intersects the y – z plane.

the crystal to be an isotropic, linear elastic medium with Young's modulus E , shear modulus μ , and Poisson ratio ν . The high shear strain region of the elastically deformed lattice (Fig. 3a) is observed at roughly 0.8 times the contact radius below the needle tip. The location of the shear strain maximum is in good agreement with hertzian theory for the strain distribution under an indenting tip¹⁷. It is also in good agreement with measurements of a two-dimensional model system formed from bubble rafts⁶. Experiments with needles of different size showed a shift of the shear strain maximum in accordance with the predictions of hertzian theory.

The role of thermal fluctuations can also be quantitatively addressed. The energy cost to create a dislocation loop of radius r is $A(\mu b^2 r/2) \ln(r/r_0)$, where r_0 is the effective core radius of the dislocation strain field, b is the length of the Burgers vector, and A is a constant that accounts for the mixed edge-screw topology of a dislocation loop^{14–16}. The energy gain from relief of the overall strain is $\pi r^2 b \tau$, where τ is the local stress before dislocation nucleation. The total nucleation energy has a maximum at $r_c = (A/4\pi)\gamma b \ln(r_c/r_0)$, while the critical energy at $r = r_c$ is $U_c = A(\mu b^2 r_c/4) \ln(r_c/r_0)$. Here, we have used $\gamma = \tau/\mu$. We calculate A by averaging over the edge and screw components of the dislocation loop to yield $A = 5/4$ (ref. 15). In addition, we include the additional energy stored within the dislocation core and use $r = b \exp(-0.4) = 0.67b$ (ref. 15). We plot the resultant r_c/b and $U_c/\mu b^3$ versus γ in Figs 4a and b. Both r_c and U_c decrease monotonically with increasing γ . The theoretical shear strength corresponds to the value of the stress at $U_c = 0$. Beyond this stress, dislocations form even in the absence of thermal fluctuations and this description breaks down. The vertical dashed line and shaded bar centred at $\gamma = 0.06$ correspond to the mean of γ and the standard deviation of the mean in the high strain region. The shaded horizontal bar centred at 6.4 marks the possible range for r_c/b determined from the experiment. The intersection of the shaded regions is close to the calculated curve, which confirms that this continuum model can be applied to our system.

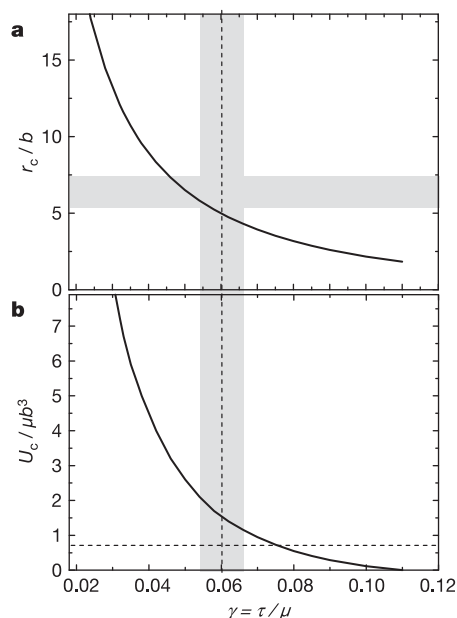


Figure 4 | Continuum model prediction for the critical radius and energy of dislocation nucleation. **a**, Dimensionless critical radius r_c/b versus shear strain for $r_0 = 0.67b$. The grey horizontal bar indicates the range of r_c/b values determined from experiments. The dashed vertical line and shaded bar centred at $\gamma = 0.06$ indicate the mean of γ and the standard deviation of the mean in the high shear strain region under the indenter. **b**, Critical nucleation energy versus shear strain. The dashed horizontal line at $U_c/\mu b^3 = 0.7$ indicates the value for the critical energy determined from experiments.

Further confirmation of the model comes from comparison of the measured and predicted values for U_c . We determine U_c from the observed nucleation rate using $J = f_0 m \exp(-U_c/kT)$, where m is the total number of particles in the high stress region and f_0 is a characteristic frequency of the particles. We estimate the nucleation rate from the time between fluctuations, $2 \times 10^{-3} \text{ s}^{-1}$. We estimate f_0 from the timescale τ required for a particle to diffuse in its local well until it reaches its nearest neighbour. The distance a particle moves is $\delta \approx 0.08 \mu\text{m}$. Because $\tau = (\pi\eta a \delta^2/kT)$ (ref. 18), where a is the particle radius, k is Boltzmann's constant, T is the temperature and the solvent viscosity is $\eta = 3 \times 10^{-3} \text{ Pa s}$ (ref. 19), we find $\tau \approx 10^{-2} \text{ s}$ yielding a frequency of $f_0 = 100 \text{ s}^{-1}$. Furthermore, we estimate $m = 125$ from the observation that the high shear strain region extends about five particle diameters in the x , y and z direction. Thus, we obtain $U_c/kT = 16$, which yields a critical nucleation energy of about 0.41 eV. Finally, using $\mu = E/[2(1 + \nu)]$ with $\nu = 1/3$ (ref. 20), $E = 0.3 \text{ Pa}$ (ref. 2), and $b = 0.94 \mu\text{m}$, we obtain $U_c/\mu b^3 = 0.7$, which is marked by the horizontal dashed line in Fig. 4b. The model predicts a value that is a factor of two larger than our estimate, which is well within the uncertainty of our estimate of $U_c/\mu b^3$.

Nano-indentation experiments on atomic systems typically entail measurements of force-displacement relationships^{3–5}. Values of forces required for indentations of atomic systems are typically several tens of micronewtons. By comparison, from the measured strain, the elastic modulus, and the contact area of the indenter, we estimate that the forces required to induce defects in these colloidal crystals are only several piconewtons and cannot be measured. We can determine the origin of this difference by estimating the shear modulus to be $\mu = U_b/a^3$, where U_b is the interaction energy. Then the difference between these forces reflects the number of particles under the tip, which leads to a factor of ~ 400 , the bond energies, which lead to a factor of ~ 50 , and the size of the particles, which leads to a factor of $\sim 10^4$; the net result gives a difference of 2×10^8 , in reasonable agreement with our estimate.

The advantage of investigating colloidal systems, however, is that we can directly image the microscopic strain field, which allows us to follow the thermal activation of dislocation loops precisely and to measure the critical size for their nucleation. The value of $U_c = 16kT$ reflects the diverging entropic cost for changing particle configurations near the close-packed limit of the colloidal crystal. By comparison, in atomic systems at equal nucleation rates, the value of U_c is a factor of two higher, reflecting that the attempt frequency is ten orders of magnitude higher. Values of μb^3 for a compliant metal such as aluminium are also only a factor of two higher than estimates for μb^3 in our colloidal suspension. Remarkably, the ratio $U_c/\mu b^3$ (Fig. 4b) is unchanged, which indicates that the effects of thermal fluctuations in our experiment are similar to those in a typical nano-indentation experiment; they should therefore be observed in experiments that have sufficiently accurate strain control.

Received 12 May; accepted 19 December 2005.

- Hirth, J. P. & Lothe, J. *Theory of Dislocations* Edn 2 (Wiley, New York, 1982).
- Schall, P., Cohen, I., Weitz, D. A. & Spaepen, F. Visualization of dislocation dynamics in colloidal crystals. *Science* **305**, 1944–1948 (2004).
- Suresh, S., Nieh, T.-G. & Choi, B. W. Nano-indentation of copper thin films on silicon substrates. *Scr. Mater.* **41**, 951–957 (1999).
- Kiely, J. D., Jarausch, K. F., Houston, J. E. & Russell, P. E. Initial stages of yield in nanoindentation. *J. Mater. Res.* **14**, 2219–2227 (1999).
- Li, J., van Vliet, K. J., Zhu, T., Yip, S. & Suresh, S. Atomistic mechanism governing elastic limit and incipient plasticity in crystals. *Nature* **418**, 307–310 (2002).
- Gouldstone, A., Van Vliet, K. J. & Suresh, S. Simulation of defect nucleation in a crystal. *Nature* **411**, 656 (2001).
- Kelchner, C. L., Plimpton, S. J. & Hamilton, J. C. Dislocation nucleation and defect structure during surface indentation. *Phys. Rev. B* **58**, 11085–11088 (1998).
- Lilleodden, E. T., Zimmerman, J. A., Foiles, S. M. & Nix, W. D. Atomistic simulations of elastic deformation and dislocation nucleation during nanoindentation. *J. Mech. Phys. Sol.* **51**, 901–920 (2003).

9. Stach, E. A. *et al.* Development of a nanoindenter for in situ transmission electron microscopy. *Microsc. Microanal.* **7**, 507–517 (2001).
10. Minor, A. M., Lilleodden, E. T., Stach, E. A. & Morris, J. W. Direct observation of incipient plasticity during nanoindentation of Al. *J. Mater. Res.* **19**, 176–182 (2004).
11. van Blaaderen, A., Ruel, R. & Wiltzius, P. Template-directed colloidal crystallization. *Nature* **385**, 321–324 (1997).
12. Weeks, E. R., Crocker, J. C., Levitt, A. C., Schofield, A. & Weitz, D. A. Three-dimensional direct space imaging of structural relaxation near the colloidal glass transition. *Science* **287**, 627–631 (2000).
13. Falk, M. L. & Langer, J. S. Dynamics of viscoplastic deformation in amorphous solids. *Phys. Rev. E* **57**, 7192–7205 (1998).
14. Frank, F. C. in *Symposium on Plastic Deformation of Crystalline Solids* 89–102 (Carnegie Institute of Technology and Office of Naval Research, Pittsburgh, 1950).
15. Hirth, J. P. in *The Relation Between the Structure and the Mechanical Properties of Metals* 217–228 (Her Majesty's Stationery Office, London, 1963).
16. Cottrell, A. H. *Dislocations and Plastic Flow in Crystals* (Oxford Univ. Press, Oxford, 1956).
17. Davies, R. M. The determination of static and dynamic yield stresses using a steel ball. *Proc. R. Soc. Lond. A* **197**, 416–432 (1949).
18. Chaikin, P. M. & Lubensky, T. C. *Principles of Condensed Matter Physics* Ch. 7.4.5 (Cambridge Univ. Press, Cambridge, 1995).
19. Higashigaki, Y., Christensen, D. H. & Wang, C. H. Studies of the reorientational motion and intermolecular interaction of dimethylsulfoxide in water by depolarized Rayleigh-scattering. *J. Phys. Chem.* **85**, 2531–2535 (1981).
20. Frenkel, D. & Ladd, A. J. C. Elastic constants of hard-sphere crystals. *Phys. Rev. Lett.* **59**, 1169 (1987).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Christianson and D. Blair for their help with the image analysis, M.L. Falk for discussions about the strain algorithm, and D. Bonner for help with programming the strain algorithm. This work was supported by a Lynen Fellowship from the Alexander von Humboldt Foundation (P.S.), and by NSF and Harvard MRSEC grants.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.S. (pschall@science.uva.nl).

LETTERS

A mechanism to thin the continental lithosphere at magma-poor margins

Luc L. Lavier¹ & Gianreto Manatschal²

Where continental plates break apart, slip along multiple normal faults provides the required space for the Earth's crust to thin and subside¹. After initial rifting, however, the displacement on normal faults observed at the sea floor seems not to match the inferred extension². Here we show that crustal thinning can be accomplished in such extensional environments by a system of conjugate concave downward faults instead of multiple normal faults. Our model predicts that these concave faults accumulate large amounts of extension and form a very thin crust (<10 km) by exhumation of mid-crustal and mantle material. This transitional crust is capped by sub-horizontal detachment surfaces over distances exceeding 100 km with little visible deformation. Our rift model is based on numerical experiments constrained by geological and geophysical observations from the Alpine Tethys and Iberia/Newfoundland margins^{3–9}. Furthermore, we suggest that the observed transition from broadly distributed and symmetric extension to localized and asymmetric rifting is directly controlled by the existence of a strong gabbroic lower crust. The presence of such lower crustal gabbros is well constrained for the Alpine Tethys system^{4,9}. Initial decoupling of upper crustal deformation from lower crustal and mantle deformation by progressive weakening of the middle crust is an essential requirement to reproduce the observed rift evolution. This is achieved in our models by the formation of weak ductile shear zones.

To allow the inclusion of mechanical decoupling of the upper crust and lower crust-mantle in numerical experiments, we propose a new formulation for the rheology of the crust (Fig. 1). Weakening associated with viscous strain accumulation has been included in previous numerical experiments and plays an important role in decoupling the upper crust from the lower crust and mantle¹⁰. However, studies in the Pogallo and Lugano–Val Grande shear zones in the Alps^{11,12} and in other ductile shear zones (DSZs)^{13,14} show that for polymineralic rocks, such as granite, localization occurs for high stress and low strain in the presence of fluids^{9,15,16}, and from low strain to high strain when the rock is transformed into a mylonite^{17,18}. Hence, softening is both stress- and strain-dependent and causes a localized decrease in viscosity with respect to the surrounding rocks.

We chose to parameterize ductile softening as though it were governed by a decrease in viscosity from plagioclase to wet quartzite rheology¹⁹ (Fig. 1c). The softening transition occurs only between the temperatures corresponding to the onset of quartz plasticity (300 °C) to that of plagioclase plasticity (450 °C). The transition occurs when the product of the effective stress σ^{eff} with the total accumulated strain $\varepsilon_t^{\text{eff}}$ is greater than $4 \times 10^6 \text{ J}$ (which is an approximate measure of the minimum work done to deform the system) (Fig. 1c) with $\sigma^{\text{eff}} = \sqrt{(\sigma^{\text{II}})}$ and $\varepsilon_t^{\text{eff}} = \sqrt{(\varepsilon_t^{\text{II}})}$ (where σ^{II} and $\varepsilon_t^{\text{II}}$ are the second invariants of the stress and strain). This was chosen to initiate an abrupt change in viscosity for stresses corresponding to the brittle–ductile transition (BDT) (100 MPa for the choice of the initial

rheology) for 4% of strain. Deformation accumulates at the BDT where the upper, elastoplastic crust shears over the viscoelastic crust, and results in the formation of sub-horizontal ductile shear zones (Fig. 1d, e). In the example depicted in Fig. 1d, the DSZ is active at low angles (5°) near the BDT and the brittle fault is active at 45°. For higher strain (>20%), the stress needed to initiate ductile softening is of the order of a few tens of MPa or less (Fig. 1e). Thus, multiple, sub-horizontal DSZs will form in the vicinity of the 300 °C isotherm where the work criterion is satisfied. Our new parameterization also causes the deformation to follow the thermal evolution in the crust. As these DSZs become highly strained, they further weaken and form a wide, viscous channel. This rheological transition is critical in controlling the evolution of rifting.

Observations of the sediment, basement and fault architecture of the Alpine Tethys and Iberia/Newfoundland margins enable a precise description of the temporal and spatial evolution of deformation during rifting^{3,4}. The pre-rift conditions were estimated from: the stratigraphic record of the Alpine Tethys, the crustal-scale seismic section across the Iberia and Newfoundland margins, and the pressure–temperature–time path through the Ivrea lower crustal section and Malenco crust-mantle boundary^{4,9}. These show that the pre-rift crust was ~30 km thick, dominated by quartzofeldspathic upper and middle crust, and with a gabbroic lower crust inherited from post-Variscan extensional collapse in the Permian period (Fig. 1a). The temperature and pressure conditions at the crust–mantle boundary were 550 °C for 0.9–1.0 GPa at the onset of rifting⁹ (Fig. 1a). The volcanic activity during rifting is lacking at both margins, arguing for a cold initial geotherm.

Using structural reconstructions, we identify three phases of rifting (Fig. 2a). The initial ‘stretching mode’, A, is characterized by distributed listric-normal faulting, cutting through the brittle upper crust and soling out at mid-crustal levels. The faults bound the rift-basins, up to 4 km deep and 30 km wide. Fault offsets are less than 10 km and total extension is limited. Both hanging wall and footwall are subsiding (A in Fig. 2a). The type examples are the Generoso basin in the Southern Alps¹² and the Jeanne d’Arc basin in Newfoundland²⁰.

The ‘thinning mode’, B, follows and affects the future distal margin in a narrow zone. This area of the margin is commonly buried under thick sediments. In seismic sections there is little evidence of upper crustal extension. Drill-holes are rare and do not provide much information about tectonic subsidence or the nature of the basement. From the Alps, pressure–temperature–time data from the Ivrea lower crustal section²¹ can be related to fault activity along a major DSZ (for example, the Pogallo shear zone²¹) while the isostatic movements are constrained by the stratigraphic record exposed in the Briançonnais and Err/Canavese domains (B in Fig. 2a). These observations suggest the occurrence of crustal-scale shear zones thinning the crust to less than 10 km, without the presence of distributed normal faulting in

¹University of Texas Institute for Geophysics, Jackson School of Geosciences, Austin, Texas 78759, USA. ²CGS-EOST, Université Louis Pasteur, 1 rue Blessig, F-67084 Strasbourg, France.

the upper crust and associated large subsidence. Such crustal-scale shear zones were described for the Iberia margin⁷ (for example, seismic reflector⁴ C), whereas more accurate geological constraints were obtained from the Alps (for example, Pogallo fault²¹) (B in Fig. 2a). Other possible examples of the thinning phase are the Galicia Interior basin near Iberia²² and the west African margin²³. In all of these examples, thinning of the crust is accompanied by little or no evidence of upper crustal deformation.

The 'exhumation mode', C, is characterized by downward concave faults that generate fault offsets (more than 10 km) without producing major observable topography, because the subcontinental mantle is exhumed at the sea floor³. Examples of this fault system were seismically imaged and drilled in the Iberia abyssal plain³ and are exposed in the Err and Platta nappes in the Alps⁴ (C in Fig. 2a). Exhumed mantle in the ocean continent transition^{5,6} is associated with late, shallow-crustal detachment systems⁷⁻⁹, and is serpentinized at temperatures under 600 °C and at depths less than 10 km, where hydrothermal circulation is occurring²⁴.

The numerical experiments were performed with an extended version of the numerical code PARAVOZ²⁵. We use initial conditions and a rheological parameterization derived from observations in the Alps (Fig. 1a). The lithosphere is modelled as a nonlinear viscoelastoplastic material²⁶. The crust is initially 30 km thick. For the first 24 km, the rheology approximates the mechanical behaviour of a polyminerale material composed of plagioclase and quartz. The deepest 6 km of the crust is gabbroic and modelled with a nonlinear Maxwell viscoelastic rheology, which is intermediate between plagioclase (anorthosite) and dry olivine. The mantle is modelled as a dry olivine. The temperature at the crust–mantle boundary is taken to be 550 °C, increasing to 1,300 °C at 100 km depth, and is set at 10 °C at the surface.

The modelled extension follows the same evolution as that predicted by geological observations. It starts by the formation of distributed normal faults rooted in the middle crust and is weakened by the formation of several DSZs (A in Fig. 2b). The weak middle crust flows over the strong lower crust and allows for the

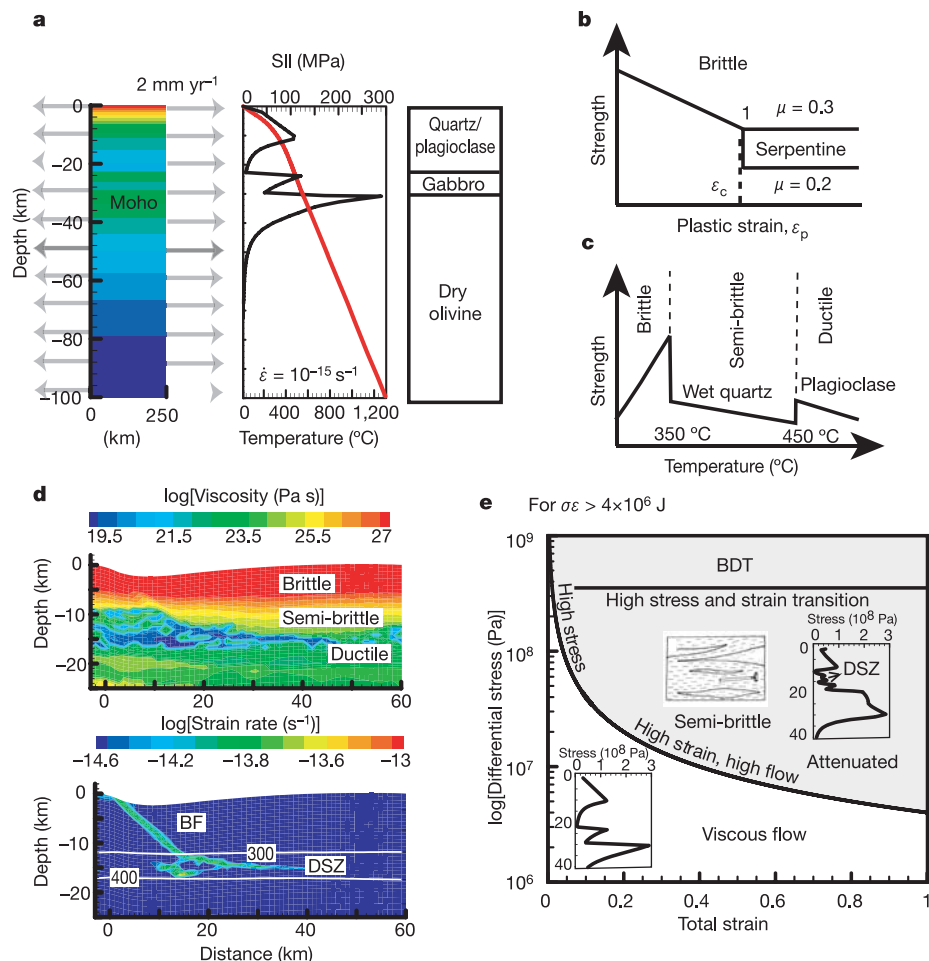


Figure 1 | Geological constraints, rheological parameterization and modelling approach. **a**, The sides of the model are pulled apart at a constant half-rate of 2 mm yr⁻¹. Isostatic equilibrium is simulated at the bottom while the top surface is stress-free to model the topographic evolution. SII is the yield stress. **b**, In the brittle areas, the lithosphere is elastoplastic with a Mohr–Coulomb yield criterion. The material is both frictional and cohesive (initially the friction coefficient $\mu = 0.6$ and the cohesion $C = 44$ MPa). Both cohesion and friction decrease locally as a function of plastic strain (10%) to trigger the formation of shear bands (to $\mu = 0.3$ and $C = 4$ MPa). Serpentinization occurs in the mantle for temperatures under 600 °C and depths under 10 km, and we further decrease the friction coefficient at this depth to 0.2 (ref. 29). Crustal and mantle density are 2,800 kg m⁻³ and 3,300 kg m⁻³ respectively. Power-law creep parameters

are: crust, quartz³⁰ (exponent $n = 3$, activation energy $Q = 2 \times 10^5$ J mol⁻¹, pre-exponent $A = 5 \times 10^2$ MPa⁻ⁿ s⁻¹) and plagioclase³¹ ($n = 3.2$, $Q = 2.38 \times 10^5$ J mol⁻¹, $A = 3.3 \times 10^{-1}$ MPa⁻ⁿ s⁻¹), gabbroic lower crust ($n = 3.05$, $Q = 3.5 \times 10^5$ J mol⁻¹, $A = 1.25 \times 10^{-1}$ MPa⁻ⁿ s⁻¹), mantle, dry olivine³² ($n = 3$, $Q = 5.2 \times 10^5$ J mol⁻¹, $A = 7 \times 10^4$ MPa⁻ⁿ s⁻¹). ϵ_c is the characteristic plastic strain for which the material has weakened. **c**, Schematic decrease in viscosity associated with the formation of a ductile shear zone. **d**, Viscosity and strain-rate fields for a listric normal fault. The brittle–plastic fault (labelled BF) roots at 12 km in a sub-horizontal low-viscosity layer (DSZ) extending down to 15 km between the 300 °C and 400 °C isotherms. **e**, Brittle to semi-brittle, semi-brittle to ductile, and brittle to ductile transitions in the stress–strain space for our parameterization.

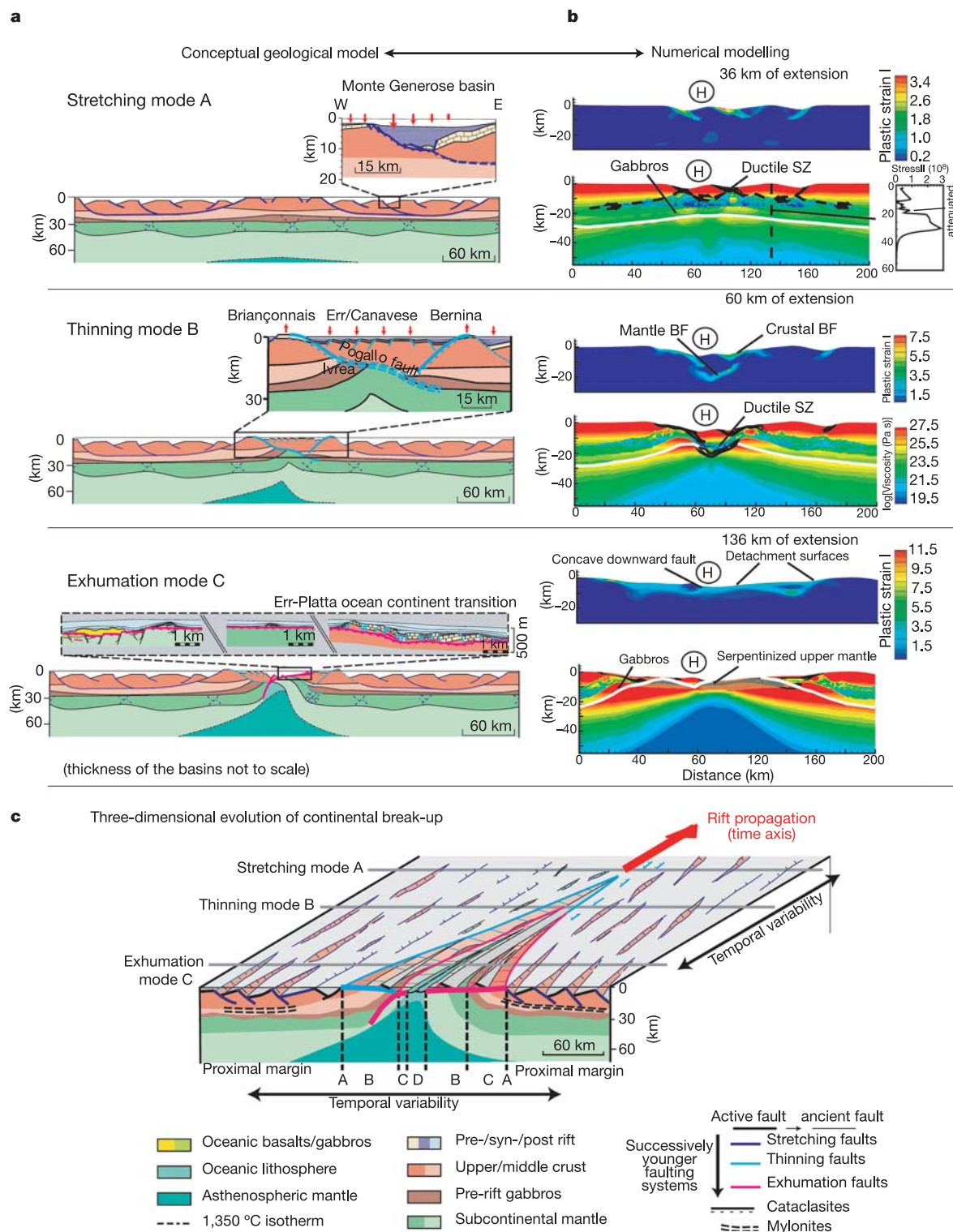


Figure 2 | Modes of extension leading to continental break-up and three-dimensional concept for the temporal and spatial evolution of rifting.

a, Conceptual models for the different phases of extension based on observations from the Alpine and Iberia/Newfoundland margins. They are set apart by different styles of deformation. The stretching mode (A) is characterized by listric faulting and a differential subsidence of half-grabens exemplified by the Monte Generoso basin¹². The thinning mode (B) is the least documented of the modes but it can be characterized by maximum thinning of the lithosphere and the presence of a major ductile shear zone (Pogallo shear zone²¹) accommodating differential motion (up to 10 km upward) between the upper crust and the lower crust/upper mantle. It is accompanied by not much uplift of the rift flanks and subsidence in the

hanging wall. The exhumation mode (C) is well documented and distinguished by the exhumation of serpentinized upper mantle from less than 10 km depth along a downward-concave fault. **b**, Modelled evolution during lithospheric extension. The plastic strain (brittle deformation) and viscosity field are plotted for the different phases of the modelled evolution of the lithosphere. Note the similarity between the observed and modelled structures. The stress envelope is plotted for a given depth profile (dashed lines for 36 km of extension). For each viscosity field the Mohorovic boundary is shown by a white line. The letter H indicates the hanging block discussed in the main text. **c**, Schematic representation of the temporal and spatial evolution of the three consecutive phases of rifting leading to break-up and seafloor spreading.

delocalization of deformation. In addition to the elastic resistance of the gabbroic lower crust, cooling of the mantle augments the mechanical strength of the lithosphere (A in Fig. 2b). This strengthening in conjunction with viscous flow in the weakened middle crust allows for the distribution of the deformation across the model. The middle crust then forms a 5–6 km decoupling, weak viscous layer composed of DSZs below the block labelled H (for hanging) in Fig. 2b. After 50 km of extension, brittle deformation in the upper mantle leads to the formation of a narrow rift zone below the hanging block H (B in Fig. 2b). Brittle normal faults in the mantle and in the upper crust are decoupled, and form a system of rolling hinge faults on both sides of block H (B in Fig. 2b). In the upper crust, normal faults accommodate large offsets (>30 km), and these thin the crust by exhuming the middle crust and shearing off both sides of block H (B in Fig. 2b). The same motion is accommodated in the mantle, resulting in the exhumation and juxtaposition of deeper mantle levels against middle-crustal rocks. It also leads to the omission of the strong lower crust and upper mantle at the surface. In this ‘thinning’ mode, the crust can be thinned to less than 10 km depth over distances of over 100 km on both sides of block H (C in Fig. 2b). Coupling of the upper crust and mantle occurs when the low

viscosity layer below block H is thinned to less than 4 km. At that point, a robust serpentinite front develops at temperatures below 600 °C (right side of block H; C in Fig. 2b) and at depths of less than 10 km. The faults in the brittle mantle and crust on the right side of block H merge to form one downward concave fault that exhumes the serpentinitized mantle at the surface (C in Fig. 2b). The resulting ‘transitional’ crust formed by exhumation of the middle crust and upper mantle spans a distance of over 100 km and is topped by detachment surfaces (C in Fig. 2b). This system sets up the structure of ocean continent transition (C in Fig. 2a).

The discovery of three consecutive phases of continental extension overprinting each other implies that, in a propagating rift system, these phases must be acting next to each other (Fig. 2c). Thus, we should see distributed listric faulting ahead of propagating rifts (stretching phase A). This distributed listric faulting is then later overprinted by localized detachment faulting that first exhumes the middle crustal levels (thinning phase B) and before exhuming the serpentinitized mantle (exhumation phase C). Final break-up then propagates across the rift sequence and separates the margins, leading to seafloor spreading. Such a progression is observed in ‘V’-shaped basins, such as the Woodlark basin, for instance²⁷.

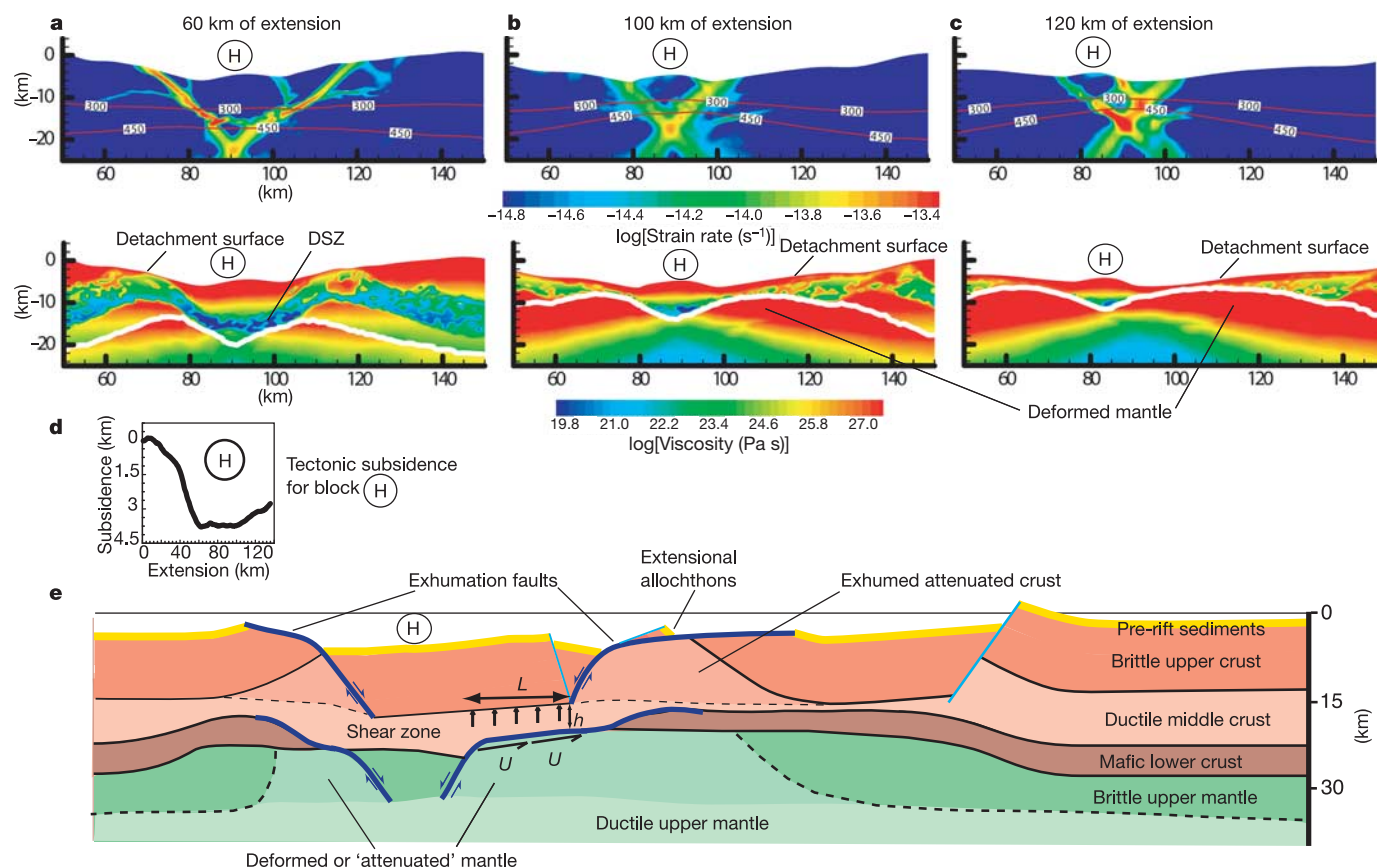


Figure 3 | Details of the modelled thinning phase and conceptual model of crustal thinning. **a–c**, Localized deformation during the thinning phase around the block labelled H. Viscous and strain-rate fields for 60 km (**a**), 100 km (**b**) and 120 km (**c**) of extension. The DSZ behaves as a lubricant between block H and the upper mantle. Block H is trapped between two active shear zones. A low-viscosity layer (between 300 °C and 450 °C) corresponds to the DSZ and is dragged at about half of the rifting rate by the motion generated by the DSZ in the mantle. Similar to the flow of a slider-bearing, flow in the DSZ is confined below and around block H. Shear in the DSZ generates a pressure change Q that can support a normal load. Using a low-Reynolds-number approximation: $Q = \eta UL/h^2$, where η is the average viscosity, U is the dragging velocity, L is the length and h is the thickness of the DSZ. For η varying between 10^{19} and 10^{21} Pa s, Q is of the order of a few

to tens of MPa. This pressure change can suppress the subsidence of block H during thinning by a few 100 m to 1 km. **d**, This effect may explain why the subsidence of block H is constant during the thinning phase (60–110 km of extension). If the stresses exerted by the DSZ are larger than the yield stress at the base of block H, they may provide the necessary work to thin the crust. This mechanism may explain the formation of extensional allochthons. Finally, the subsidence curve shows that block H is uplifted during the final phase of break-up. This can be explained by the upward-restoring force, resulting in deep necking below the Mohorovic depth in the last phase of rifting³³. **e**, Conceptual model for the thinning phase during continental break-up. We note that thinning is accommodated by the simultaneous exhumation of middle crust and upper mantle.

Our modelling highlights a new mechanism for lithospheric thinning that can reproduce the major observations made at the Iberia/Newfoundland and Alpine Tethys margins, and evolves from distributed and decoupled stretching phase to a final, localized and coupled, exhumation phase. We find two processes that are likely to weaken the lithosphere and may explain the evolution of rifting: (1) attenuation of the middle crust in the initial stage of rifting, and, (2) serpentinization of the lithosphere during the last exhumation phase. Serpentinization is a critical mechanism to the establishment of a concave-downward fault, because it allows for the weakening of the lithosphere as it bends upward through the rolling hinge. It also replaces the formation of sub-horizontal ductile shear zones as the mechanism that weakens the lithosphere during exhumation. Mid-crustal weakening and serpentinization allow for thinning and exhumation of an originally strong mantle. The presence of an inherited gabbroic lower crust initially leads to distributed thinning in the presence of a weakened middle crust.

The model suggests that exhumation of middle crustal rocks occurs simultaneously with lower crust/mantle thinning along concave, downward faults (Fig. 3 legend). Exhumation leads to the formation of new tectonic surfaces (detachment surfaces) and the lack of rifted crust or pre-rift sediment in the area of continental break-up. This leaves us with no 'seismically detectable' indicator of the amount of strain incurred during thinning. This process may explain poorly understood observations of tectonic subsidence in the absence of observable deformation in the upper crust at other margins (for example, the South Atlantic margin^{2–23}). The inferred rheological evolution of the lithosphere shows that weakening is acting to minimize the force that is needed to bend the crust and upper mantle by decoupling them throughout the rift history²⁸. In the absence of magmatic activity to weaken the lithosphere, the numerical experiments point to these two processes as key mechanisms allowing continental break-up in a strong lithosphere.

Received 13 October 2005; accepted 24 January 2006.

- Vening-Meisnez, F. A. Les grabens Africains résultants de compression ou de tension de la croûte terrestre? *Mém. Inst. R. Colon. Belge* **21**, 539–552 (1950).
- Karner, G. D., Driscoll, N. W. & Barker, D. H. N. in *Petroleum Systems and Evolving Technologies in African Exploration and Production* (eds Arthur, T., MacGregor, D. & Cameron, N. R.). *Spec. Publ. Geol. Soc. Lond.* **207**, 105–129 (2003).
- Whitmarsh, R. B., Manatschal, G. & Minshull, T. A. Evolution of magma-poor continental margins from rifting to sea-floor spreading. *Nature* **413**, 150–154 (2001).
- Manatschal, G. New models for evolution of magma-poor rifted margins based on a review of data and concepts from West Iberia and the Alps. *Int. J. Earth Sci.* **93**, 432–466 (2004).
- Decandia, F. A. & La Elter, P. "Zona" ofiolitífera del Bracco nel settore compreso fra Levante e la Val Graveglia (Appennino ligure). *Mem. Soc. Geol. Ital.* **11**, 503–530 (1972).
- Boillot, G. *et al.* Tectonic denudation of the upper mantle along passive margins: A model based on drilling results (ODP Leg 103, western Galicia margin, Spain). *Tectonophysics* **132**, 335–342 (1987).
- Froitzheim, N. & Eberli, G. P. Extensional detachment faulting in the evolution of a Tethys passive continental margin, eastern Alps, Switzerland. *Geol. Soc. Am. Bull.* **102**, 1297–1308 (1990).
- Reston, T. J., Krawczyk, C. M. & Hoffmann, H. J. in *The Tectonics, Sedimentation and Palaeoceanography of the North Atlantic Region* (eds Scrutton, R. A., Stoker, M. S., Shimmield, G. B. & Tudhope, A. W.) *Geol. Soc. Spec. Publ.* **90**, 93–109 (1995).
- Müntener, O., Hermann, J. & Trommsdorff, V. Cooling history and exhumation of lower-crustal granulite and upper mantle (Malenco, eastern central Alps). *J. Petrol.* **41**, 175–200 (2000).
- Huisman, R. & Beaumont, C. H. Symmetric and asymmetric lithospheric extension; relative effects of frictional-plastic and viscous strain softening. *J. Geophys. Res.* **108**, 10.1029/2002JB002026 (2003).
- Handy, M. R. Deformation regimes and the rheological evolution of fault zones in the lithosphere: the effects of pressure, temperature, grain size and time. *Tectonophysics* **163**, 119–152 (1989).
- Bertotti, G. Early Mesozoic extension and Alpine shortening in the western Southern Alps: The geology of the area between Lugano and Menaggio (Lombardy, northern Italy). *Mem. Sci. Geol. (Padova)* **43**, 17–123 (1991).
- Rutter, E. H., Brodie, K. H. & Evans, P. J. in *The Geometry of Naturally Deformed Rocks* (eds Casey, M., Dietrich, D., Ford, M., Watkinson, J. & Hudleston, P. J.) *J. Struct. Geol.* **15**, 647–662 (1993).
- Hirth, G. & Tullis, J. Dislocation creep regimes in quartz aggregates. *J. Struct. Geol.* **14**, 145–159 (1992).
- Rutter, E. H. & Brodie, K. H. in *Continental Lower Crust* (eds Fountain, D. M., Arculus, R. J. & Kay, R. W.) 201–267 (Elsevier, New York, 1992).
- Fricke, H. C., Wickham, S. M. & O'Neil, J. R. Oxygen and hydrogen isotope evidence for meteoric water infiltration during mylonitization and uplift in the Ruby Mountains-East Humboldt Range core complex, Nevada. *Contrib. Mineral. Petrol.* **111**, 203–221 (1992).
- White, S. H., Burrows, S. E., Carreras, J., Shaw, N. D. & Humphreys, F. J. On mylonites in ductile shear zones. *J. Struct. Geol.* **2**, 175–187 (1980).
- Jordan, P. The rheology of polyminerals rocks—an approach. *Geol. Rundsch.* **77**, 285–294 (1988).
- Gilbert, L. E., Scholz, C. H. & Beavan, J. Strain localization along the San Andreas Fault; consequences for loading mechanisms. *J. Geophys. Res.* **99**, 23975–23984 (1994).
- Tankard, A. J., Welsink, H. J. & Jenkins, W. A. M. in *Extensional Tectonics and Stratigraphy of the North Atlantic Margins* (eds Tankard, A. J. & Balkwill, H. R.) *Am. Assoc. Petroleum Geol. Mem.* **46**, 265–282 (1989).
- Handy, M. R. & Zingg, A. The tectonic and rheological evolution of an attenuated cross section of the continental crust: Ivrea crustal section, southern Alps, northwestern Italy and southern Switzerland. *Geol. Soc. Am. Bull.* **103**, 236–253 (1991).
- Pérez-Gussinyé, M., Ranero, C. R., Reston, T. J. & Sawyer, D. Mechanisms of extension at non-volcanic margins: Evidence from the Galicia interior basin, west of Iberia. *J. Geophys. Res.* **108**, 10.1029/2001JB000901 (2003).
- Contrucci, I. *et al.* Deep structure of the West African continental margin (Congo, Zaïre, Angola), between 5°S and 8°S, from reflection/refraction seismics and gravity data. *Geophys. J. Int.* **158**, 529–553 (2004).
- Ingebritsen, S. E. & Manning, C. E. Geological implications of a permeability-depth curve for the continental crust. *Geology* **27**, 1107–1110 (1999).
- Poliakov, A. N. B., Podladchikov, Y. & Talbot, C. Initiation of salt diapirs with frictional overburdens—numerical experiments. *Tectonophysics* **228**, 199–210 (1993).
- Taylor, B., Goodliffe, A. M. & Martinez, M. How continents break up: Insights from Papua New Guinea. *J. Geophys. Res.* **104**, 7497–7512 (1999).
- McNutt, M. K., Diamant, M. & Kogan, M. G. Variations of elastic plate thickness at continental thrust belts. *J. Geophys. Res.* **93**, 8825–8838 (1988).
- Lavier, L. L. & Buck, W. R. Half-graben vs. large-offset low-angle normal fault: The importance of keeping cool during normal faulting. *J. Geophys. Res.* **107**, doi:10.1029/2001JB000513 (2002).
- Escartin, J., Hirth, G. & Evans, B. Effects of serpentinization on the lithospheric strength and the style of normal faulting at slow-spreading ridges. *Earth Planet. Sci. Lett.* **151**, 181–190 (1997).
- Brace, W. F. & Kohlstedt, D. L. Limit on lithospheric stress imposed by laboratory experiments. *J. Geophys. Res.* **85**, 6248–6252 (1980).
- Shelton, G. & Tullis, J. A. Experimental flow laws for crustal rocks. *Trans. Am. Geophys. Union* **62**, 396 (1981).
- Goetze, C. The mechanisms of creep in olivine. *Phil. Trans. R. Soc. Lond.* **288**, 99–119 (1978).
- Braun, J. & Beaumont, C. in *Sedimentary Basins and Basin-Forming Mechanisms* (eds Beaumont, C. & Tankard, A.) *Can. Soc. Petrol. Geol. Mem.* **12**, 241–258 (1987).

Acknowledgements We thank O. Müntener, G. Péron-Pinvidic, H. Van Avendonk, W. Powell and N. Bangs for help with preparation of the manuscript. We thank T. J. Reston for comments on the manuscript. G.M. was funded by a grant from the programme 'GDR Marge'. L.L.L. was supported by a grant from ExxonMobil Upstream Research Company.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.L.L. (luc@ig.utexas.edu).

A new carnivorous dinosaur from the Late Jurassic Solnhofen archipelago

Ursula B. Göhlich¹ & Luis M. Chiappe²

Small Late Jurassic theropod dinosaurs are rare worldwide. In Europe these carnivorous dinosaurs are represented primarily by only two skeletons of *Compsognathus*^{1,2}, neither of which is well preserved. Here we describe a small new theropod dinosaur from the Late Jurassic period of Schamhaupten in southern Germany^{3,4}. Being exquisitely preserved and complete from the snout to the distal third of the tail, the new fossil is the best-preserved predatory, non-avian dinosaur in Europe. It possesses a suite of characters that support its identification as a basal coelurosaur. A cladistic analysis indicates that the new taxon is closer to maniraptorans than to tyrannosauroids, grouping it with taxa often considered to be compsognathids. Large portions of integument are preserved along its tail. The absence of feathers or feather-like structures in a fossil phylogenetically nested within feathered theropods^{5,6} indicates that the evolution of these integumentary structures might be more complex than previously thought.

The new Schamhaupten fossil is the second non-avian theropod found in the laminated limestones of the Late Jurassic Solnhofen reef archipelago of Bavaria⁷, after the discovery of the celebrated *Compsognathus* nearly 150 years ago. Biostratigraphic studies of the Schamhaupten limestones indicate that they are 151–152 million years old^{3,4,8}, and are therefore slightly older than those containing *Compsognathus* and the famous bird *Archaeopteryx*.

Dinosauria Owen, 1842
Theropoda Marsh, 1881C
Tetanurae Gauthier, 1986
Coelurosauria Huene, 1914
Compsognathidae Marsh, 1882
Juravenator starki gen. et sp. nov.

Etymology. *Jura*, referring to the Bavarian Jura mountains; plus *venator* (Latin), a hunter. The species name, *starki*, honours the family Stark, owner of the Quarry Stark.

Holotype. JME Sch 200 (Jura-Museum Eichstätt), a nearly complete and articulated skeleton missing only the distal third of its tail (Fig. 1). The strong scarring and pitting of the bone surface⁹, the lack of fusion between sacral vertebrae, and the presence of open neurocentral sutures¹⁰ demonstrate that JME Sch 200 is a juvenile.

Horizon and locality. Silicified, laminated limestone, Late Jurassic, Upper Malm, Late Kimmeridgian, Malm Epsilon 2 (*setatum*-subzone), Quarry Stark, west of Schamhaupten, Southern Franconian Alb, Bavaria, Germany (see Supplementary Information).

Diagnosis. Small basal coelurosaur (preserved length 65 cm; estimated total length 75–80 cm; see Supplementary Information for measurements) with a small number (eight) of maxillary teeth, no premaxillary–maxillary diastema, posterior serrations on premaxillary teeth, concave rostral margin of the jugal process of the

postorbital, relatively long scapula with narrowest portion at neck, and proportionally short feet. *Juravenator starki* may also be diagnosed by the presence of an antorbital fenestra subequal in length to the orbit and by an abbreviated deltopectoral crest of the humerus, although these features may be related to the early ontogenetic age of the holotype. In addition, *Juravenator* is unique among basal coelurosaurs in having proximally high manual claws that taper abruptly at midpoint, a ventrally notched premaxillary–maxillary contact, and bow-like zygapophyses in mid-caudal vertebrae; these features are interpreted as autapomorphies.

Description and comparisons. The skull of JME Sch 200 is proportionally large (Fig. 1). It has a moderately long rostrum, elliptical external nares bordered by the premaxilla and nasal, a big antorbital fossa largely perforated by an antorbital fenestra and a small maxillary fenestra, round orbits, a flat cranial roof, and teeth lacking anterior serrations (Fig. 2). The premaxilla bears three, perhaps four, teeth—the penultimate one posteriorly serrated (Fig. 1). The maxillary teeth are more recurved than the premaxillary ones. The lacrimal has the shape of an inverted 'L'. The jugal is slender with a well-developed postorbital ramus and a tapering subtemporal ramus that ends shortly behind the base of the latter. The frontal is large and it participates extensively of the supratemporal fenestra. The parietal is one-third to one-quarter the length of the frontal. The frontal ramus of the T-shaped postorbital is slightly longer than the squamosal ramus and shorter than the jugal ramus. The lower jaw of *Juravenator* lacks a mandibular foramen, as in the compsognathids¹¹ *Compsognathus*¹, *Sinosauropteryx*¹², *Huaxiagnathus*¹³ and *Scipionyx*¹⁴ and other basal coelurosaurs such as tyrannosauroids, and its teeth—probably not more than 11—extend caudally to nearly the end of the upper dentition.

There are eight to ten cervical vertebrae. These have short and low neural spines and, as in compsognathids^{11,12,13}, extremely long, hair-like ribs (Fig. 1). The trunk is composed of 13 dorsal vertebrae with distally expanded neural spines, another synapomorphy of compsognathids^{11–13,15}. The tail is extremely long—comparisons with *Sinosauropteryx* indicate that the 44 caudal vertebrae preserved in JME Sch 200 represent about two-thirds of the complete tail. Such an elongated tail is comparable to the tail of *Sinosauropteryx* and longer than in most other theropods¹⁶. The caudal centra are elongated and very similar throughout the tail. Their length remains more or less constant until the 17th caudal, when the length of the centra starts increasing. The transition point is around the 14th to 15th vertebra. The short zygapophyses of the middle caudals have a uniquely hooked morphology (Fig. 1). The chevrons are elongate and rod-like shaped; these bones decrease in length gradually towards the distal end of the tail.

The scapula is slender and elongate (its length is about ten times its width at mid-shaft and 80% of the length of the femur). This bone has a prominent and subtriangular acromion and a scapular blade

¹Department for Geo- and Environmental Sciences, Section Paleontology, University of Munich, Richard-Wagner-Strasse 10, D-80333 Munich, Germany. ²The Dinosaur Institute, Natural History Museum of Los Angeles County, 900 Exposition Boulevard, Los Angeles, California 90007, USA.

that gradually expands towards its tip. The glenoid is rimmed and directed caudally. The forelimb is 50% of the length of the hindlimb (measured along their longest digit and including their corresponding claws)—*Juravenator* has the longest forelimb among compsognathids. The deltopectoral crest of the humerus is proximodistally very short and triangular in shape. The radius and ulna are straight and about two-thirds the length of the humerus. Metacarpal I is about 40% of the length of metacarpal II but of about the same width. Metacarpal III is about half the width of metacarpal II and 75% of its length. The hand of *Juravenator* carries three robust and clawed digits (Fig. 1) with two (I), three (II) and four (III) phalanges. Digit I is the shortest and digit II the longest. The pelvis consists of a low and straight ilium with rounded and squared-off anterior and posterior ends, respectively. The other elements of the pelvis are of difficult interpretation. The hindlimbs are robust. The femur is 10% shorter than the tibia, which is similarly shorter than the foot (metatarsal III and digit including claw). Metatarsal III is the longest, followed by metatarsals IV and II. Likewise, pedal digit III is the longest, followed by digits IV, II and I; the phalangeal formula of these digits is 2-3-4-5.

Soft tissue is preserved along the tibiae, and particularly between the 8th and the 22nd caudal vertebrae, where it defines the outline of the tail. The latter section allows observation of the skin surface and other soft parts (Fig. 3). The integument of *Juravenator* is formed of uniformly sized, smooth tubercles (about 15 tubercles per 25 mm² of preserved tissue) similar in appearance to the small, conical and

non-imbricated tubercles of many other non-avian dinosaurs^{17,18}. An array of feathers has been discovered among non-avian coelurosaurs⁶. However, the absence of either feathers or skin follicles associated with the preserved integument of *Juravenator* indicates that at least the central portion of the tail of this coelurosaur was devoid of plumage. The remaining soft tissue is represented by a series of fibres ventral to the haemal arches of the 10th to 14th caudals and parallel to the axis of the tail. These fibres probably represent tendons of the hypaxial musculature and ligaments of the tail¹⁹, as interpreted for similar soft parts associated with the skeleton of *Scipionyx*¹⁴, although they could also correspond to bundles of subcutaneous collagen fibres²⁰.

A cladistic analysis of 35 theropod taxa (Supplementary Information) clusters *Juravenator* with a diversity of other basal, small-bodied coelurosaurs of Late Jurassic to Early Cretaceous age. When the analysis is conducted with the exclusion of largely incomplete taxa, *Juravenator* nests together with compsognathids such as *Compsognathus*, *Sinosauroptryx* and *Huaxiagnathus* (Fig. 4). This condensed analysis recovers a monophyletic Compsognathidae and supports the placement of *Juravenator* within this clade.

The discovery of feathery integumentary coverings in a variety of coelurosaurs (for example, tyrannosauroids, compsognathids, therizinosauroids, oviraptorosaurs, alvarezsaurids and dromaeosaurids) has cemented the notion that feathers are a synapomorphy of this entire clade^{5,6}. Thus, the absence of feathers in *Juravenator*, a

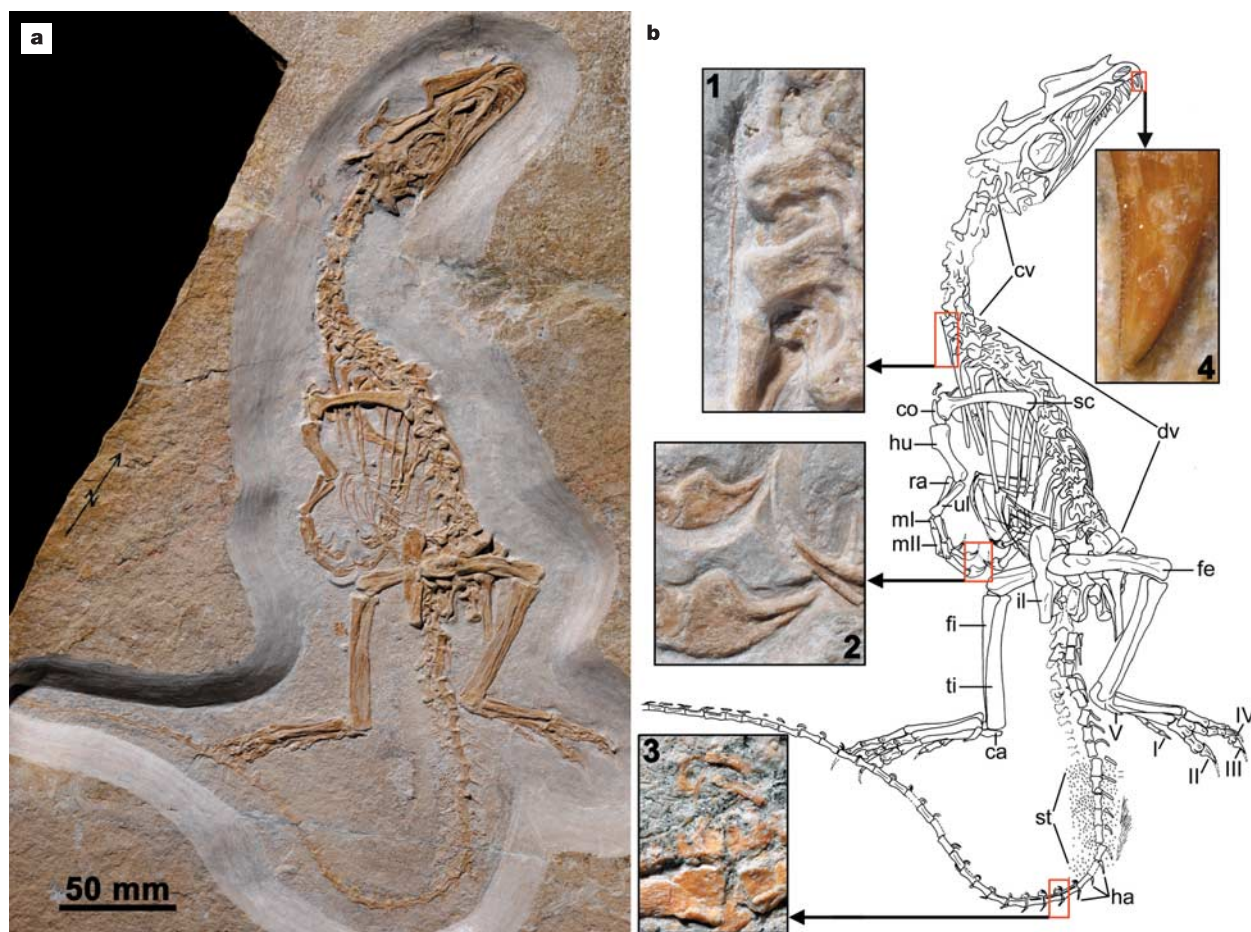


Figure 1 | Holotype of *Juravenator starki*. **a**, Specimen photographed under normal light. **b**, The presence of 'hair-like' cervical ribs (inset 1) is synapomorphic of Compsognathidae. The proximally high manual claws that taper abruptly at their midpoint (inset 2) and bow-like zygapophyses of mid-caudal vertebrae (inset 3) are regarded as autapomorphies of *Juravenator starki*. The serrated premaxillary teeth (inset 4) also distinguish

this taxon from most other basal coelurosaurs. Abbreviations: ca, calcaneus; co, coracoid; cv, cervical vertebrae; dv, dorsal vertebrae; fe, femur; fi, fibula; ha, haemal arches; hu, humerus; il, ilium; mI, metacarpal I; mII, metacarpal II; ra, radius; sc, scapula; st, soft tissue; ti, tibia; ul, ulna; I–IV, pedal digits I–IV; V, metatarsal V.

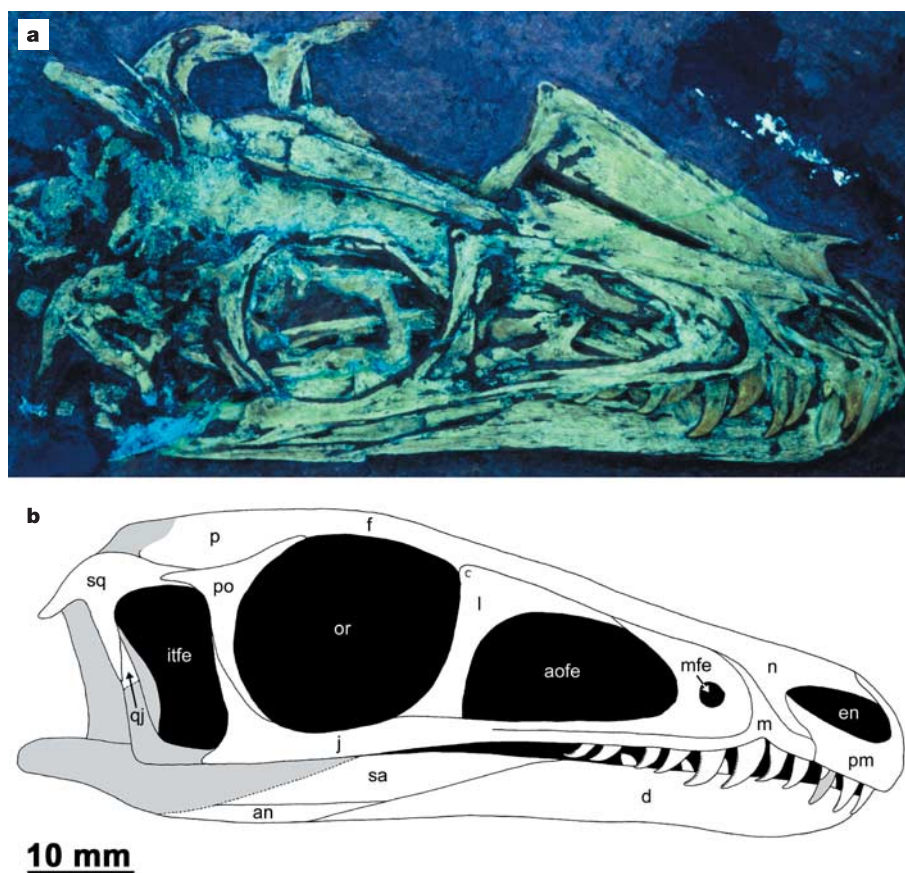


Figure 2 | Skull and mandible of *Juravenator starki*. **a**, Specimen photographed under ultraviolet light. **b**, As reconstructed (shaded areas show missing portions). Abbreviations: an, angular; aofe, antorbital fenestra; d, dentary; en, external naris; f, frontal; itfe, infratemporal fenestra;

j, jugal; l, lacrimal; m, maxilla; mfe, maxillary fenestra; n, nasal; or, orbit; p, parietal; pm, premaxilla; po, postorbital; qj, quadratojugal; sq, squamosal; sa, surangular.

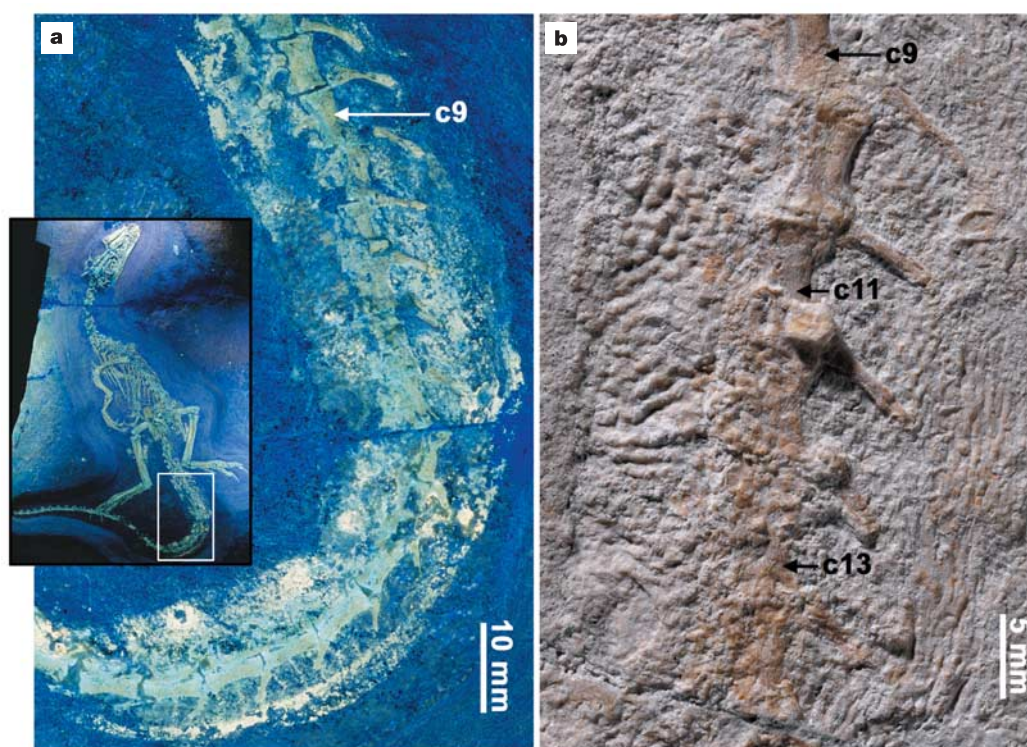


Figure 3 | Integument of *Juravenator starki*. **a**, Specimen photographed under ultraviolet light. **b**, Specimen photographed under normal light. Abbreviations: c9, c11 and c13, caudal vertebrae 9, 11 and 13.

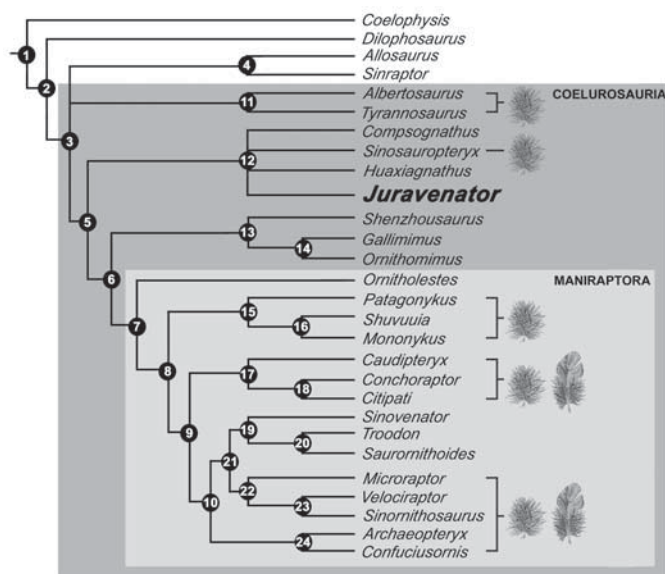


Figure 4 | Strict consensus cladogram. The eight most parsimonious trees (length 497, consistency index 0.45, retention index 0.65) for 28 theropod taxa, including *Juravenator starki* and 189 variables, are shown (see Supplementary Information). In spite of lacking feathers in the preserved integumentary portions, *Juravenator starki* is grouped together with coelurosaur clades known for having feathery coverings⁶. Plumulaceous and/or pennaceous feathers have been discovered in taxa assigned to Tyrannosauroidae (namely *Dilong*²¹; in this cladogram, tyrannosauroids are represented by the more advanced *Tyrannosaurus* and *Albertosaurus*), Compsognathidae (namely *Sinosauropteryx*^{12,16}), Alvarezsauridae (namely *Shuvuuia*²³), Oviraptorosauria (namely *Caudipteryx*²⁴), Dromaeosauridae (namely *Microraptor*²⁵ and *Sinornithosaurus*²⁶) and Aves (namely *Archaeopteryx* and *Confuciusornis*)²⁸.

taxon otherwise nested within feathered coelurosaurs, is noteworthy (Fig. 4). The extent to which feathers covered the body of these non-avian dinosaurs is not well known for some taxa (for example, the tyrannosauroid *Dilong*²¹, the therizinosauroid *Beipiaosaurus*²², and the alvarezsaurid *Shuvuuia*^{6,23}) but complete specimens of *Sinosauropteryx*^{12,16}, the oviraptorosaur *Caudipteryx*²³, and several dromaeosaurids^{24–26} indicate that the body of these animals was for the most part feathered. The fact that *Juravenator* lacks any evidence of feathers in portions of integument otherwise feathered in these coelurosaurs indicates that these animals may have differed greatly in the extension of their feathery covering. However, the role of ontogeny and seasonality in the development of the plumage of these dinosaurs remains uncertain, and the possibility cannot be ruled out that feathers evolved more than once or became lost in taxa such as *Juravenator* (Fig. 4).

The exquisitely preserved skeleton of *Juravenator starki* is the most complete non-avian theropod so far discovered in Europe, adding to the diversity of dinosaurs from this continent and in particular to the meagre record of Late Jurassic small theropods²⁷. Its discovery sheds light on a poorly known segment of coelurosaur history and indicates that the evolution of feathers in theropods might have been more complex than previously envisaged.

Received 1 September 2005; accepted 10 January 2006.

- Ostrom, J. H. The osteology of *Compsognathus longipes* Wagner. *Zitteliana* 4, 73–118 (1978).
- Bidar, A., Demay, L. & Thorne, G. *Compsognathus corallestris*, nouvelle espèce de dinosaurien théropode du Portlandien de Canjuers (Sud-Est de la France). *Extr. Ann. Mus. Hist. Natur. Nice* 1, 3–34 (1972).

- Viohl, G. & Zapp, M. Schamhaupten, an outstanding fossil-lagerstätte in a silicified plattenkalk (Kimmeridgian–Tithonian Boundary, Southern Franconian Alb, Bavaria). *Zitteliana B* 26, 27 (2005).
- Renesto, S. & Viohl, G. A sphenodontid (Reptilia, Diapsida) from the Late Kimmeridgian of Schamhaupten (Southern Franconian Alb, Bavaria, Germany). *Archaeopteryx* 15, 27–46 (1997).
- Chiappe, L. M. & Dyke, G. J. The Mesozoic radiation of birds. *Annu. Rev. Ecol. Syst.* 33, 91–124 (2002).
- Norell, M. A. & Xu, X. Feathered dinosaurs. *Annu. Rev. Earth Planet. Sci.* 33, 277–299 (2005).
- Barthel, K. W., Swinburne, N. H. M. & Conway Morris, S. *Solnhofen—A Study in Mesozoic Palaeontology* (Cambridge Univ. Press, Cambridge, 1990).
- Viohl, G. Fund eines neuen kleinen Theropoden. *Archaeopteryx* 17, 15–19 (1999).
- Sanz, J. L. et al. A nestling bird from the Lower Cretaceous of Spain: implications for avian skull and neck evolution. *Science* 276, 1543–1546 (1997).
- Brochu, C. A. Closure of neurocentral sutures during crocodilian ontogeny: implications for maturity assessment in fossil archosaurs. *J. Vertebr. Paleontol.* 16, 49–62 (1996).
- Peyer, K. A reevaluation of the French *Compsognathus* of the Tithonian of South-eastern France and its Phylogenetic Relationship with Other *Compsognathids* and *Coelurosaurs* in General. Thesis, Muséum National d'Histoire Naturelle, Paris (2004).
- Currie, P. J. & Chen, P.-J. Anatomy of *Sinosauropteryx prima* from Liaoning, northeastern China. *Can. J. Earth Sci.* 38, 1705–1727 (2001).
- Hwang, S. H., Norell, M. A., Qiang, J. & Qeqin, G. A large compsognathid from the Early Cretaceous Yixian Formation of China. *J. Syst. Paleontol.* 2, 13–30 (2004).
- Dal Sasso, C. & Signore, M. Exceptional soft-tissue preservation in a theropod dinosaur from Italy. *Nature* 392, 383–387 (1998).
- Naish, D., Martill, D. M. & Frey, E. Ecology, systematics and biogeographical relationships of dinosaurs, including a new theropod, from the Santana Formation (?Albian, Early Cretaceous) of Brazil. *Hist. Biol.* 16, 57–70 (2004).
- Chen, P.-J., Dong, Z.-M. & Zhen, S.-N. An exceptionally well-preserved theropod dinosaur from the Yixian Formation of China. *Nature* 391, 147–152 (1998).
- Czerkas, S. in *Encyclopedia of Dinosaurs* (eds Currie, P. J. & Padian, K.) 669–675 (Academic, San Diego, 1997).
- Chiappe, L. M. et al. Sauropod dinosaur embryos from the Late Cretaceous of Patagonia. *Nature* 396, 258–261 (1998).
- Frey, E. Anatomie des Körperstammes von *Alligator mississippiensis* Daudin. *Stuttg. Beitr. Naturkd. A* 424, 1–106 (1988).
- Lingham-Soliar, T. Evolution of birds: ichthyosaur integumental fibers conform to dromaeosaur protofeathers. *Naturwissenschaften* 90, 428–432 (2003).
- Xu, X. et al. Basal tyrannosauroids from China and evidence for protofeathers in tyrannosauroids. *Nature* 431, 680–684 (2004).
- Xu, X., Tang, Z.-L. & Wang, X.-L. A therizinosauroid dinosaur with integumentary structures from China. *Nature* 399, 350–354 (1999).
- Schweitzer, M. H. et al. Beta-keratin specific immunological reactivity in feather-like structures of the Cretaceous alvarezsaurid, *Shuvuuia deserti*. *J. Exp. Zool.* 285, 146–157 (1999).
- Ji, Q., Currie, P. J., Norell, M. A. & Ji, S.-A. Two feathered dinosaurs from northeastern China. *Nature* 393, 753–761 (1998).
- Xu, X., Wang, X.-L. & Wu, X.-C. A dromaeosaurid dinosaur with a filamentous integument from the Yixian Formation of China. *Nature* 401, 262–266 (1999).
- Xu, X. et al. Four-winged dinosaurs from China. *Nature* 421, 335–340 (2003).
- Weishampel, D. B., Dodson, P. & Osmólska, H. *The Dinosauria* 2nd edn (Univ. of California Press, Berkeley, 2004).
- Chiappe, L. M. & Witmer, L. M. *Mesozoic Birds: Above the Heads of Dinosaurs* (Univ. of California Press, Berkeley, 2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Weiss and K.-D. Weiss for discovery of the specimen, P. Völkl for its preparation, H. Tischlinger and G. Janssen for photography, M. Kölbl-Ebert and F. Stark for access to the specimen, and E. Frey, D. Goujet, S. Hwang, R. Leinfelder, O. Rauhut, M. Röper, G. Viohl, P. Wellnhofer and W. Werner for providing access to specimens, for logistics, and/or for discussions. This research was made possible by the Humboldt Foundation, and was also supported by the Jurassic Foundation, European Commission's Research Infrastructure Action-Synthesys Program, Antorchas Foundation, Bavarian State Collection for Paleontology and Geology, Department for Geo- and Environmental Sciences (Munich University), DFG, Jura Museum, and Natural History Museum of Los Angeles County.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.M.C. (chiappe@nhm.org).

Ultrasonic communication in frogs

Albert S. Feng^{1*}, Peter M. Narins^{2*}, Chun-He Xu³, Wen-Yu Lin¹, Zu-Lin Yu⁴, Qiang Qiu⁴, Zhi-Min Xu⁴
& Jun-Xian Shen^{4*}

Among vertebrates, only microchiropteran bats, cetaceans and some rodents are known to produce and detect ultrasounds (frequencies greater than 20 kHz) for the purpose of communication and/or echolocation, suggesting that this capacity might be restricted to mammals^{1,2}. Amphibians, reptiles and most birds generally have limited hearing capacity, with the ability to detect and produce sounds below ~12 kHz. Here we report evidence of ultrasonic communication in an amphibian, the concave-eared torrent frog (*Amolops tormotus*) from Huangshan Hot Springs, China. Males of *A. tormotus* produce diverse bird-like melodic calls with pronounced frequency modulations that often contain spectral energy in the ultrasonic range^{3,4}. To determine whether *A. tormotus* communicates using ultrasound to avoid masking by the wideband background noise of local fast-flowing streams, or whether the ultrasound is simply a by-product of the sound-production mechanism, we conducted acoustic playback experiments in the frogs' natural habitat. We found that the audible as well as the ultrasonic components of an *A. tormotus* call can evoke male vocal responses. Electrophysiological recordings from the auditory midbrain confirmed the ultrasonic hearing capacity of these frogs and that of a sympatric species facing similar environmental constraints. This extraordinary upward extension into the ultrasonic range of both the harmonic content of the advertisement calls and the frog's hearing sensitivity is likely to have co-evolved in response to the intense, predominantly low-frequency ambient noise from local streams. Because amphibians are a distinct evolutionary lineage from microchiropterans and cetaceans (which have evolved ultrasonic hearing to minimize congestion in the frequency bands used for sound communication⁵ and to increase hunting efficacy in darkness²), ultrasonic perception in these animals represents a new example of independent evolution.

We recorded the vocalization patterns of eight male frogs in the field under three experimental conditions for a period of three minutes each: (1) an NS period, during which no sound was presented, (2) a US period, during which we presented the ultrasonic components of a previously-recorded conspecific vocal signal at ~77 dB sound pressure level (SPL) (root mean squared or r.m.s. reading), a sound level that is behaviourally relevant, and (3) an AUD period, during which we presented the audible components (<20 kHz) of the same vocal signal at a similar sound level. For five frogs (asterisks in Fig. 1a), the male's calling rate was markedly increased during the AUD and/or US period, compared to spontaneous calling rates during the NS period. Three frogs (601-4, 602-1, 602-2) showed no overt evoked vocal responses to any playback stimulus. The stimulatory effect of the US components was most robust for frogs 531-1 and 601-2. Frog 531-1 did not produce any calls during the NS period, but emitted 11 calls during the US period. Frog 601-2 produced six calls during the NS period, and emitted 18

calls during the US period, including four antiphonal responses that were precisely time-locked (within 30 ms of the stimulus offset) to the US stimulus (Fig. 1b)—the probability of all four occurrences by chance is 7.4×10^{-7} (binomial probability). These results show that males of *A. tormotus* detect and respond to ultrasound.

a Evoked vocal responses

Frog no.	NS	US	AUD
*531 - 1	0	11	10 (2)
*531 - 2	2	6	—
*601 - 2	6	18 (4)	—
*531 - 3	0	0	18
*601 - 5	6	6	14
601 - 4	0	1	1 (1)
602 - 1	3	5 (1)	1
602 - 2	0	0	1

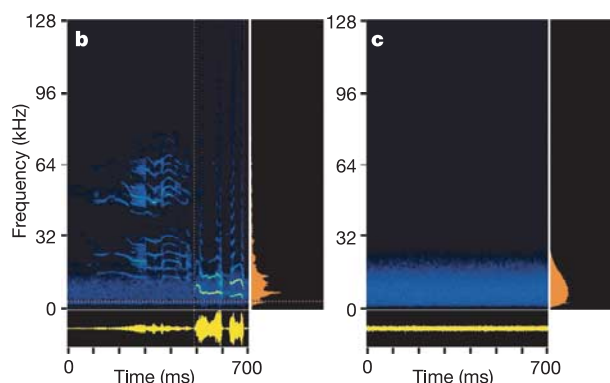


Figure 1 | *A. tormotus* can detect and respond to ultrasound. **a**, Evoked vocal responses from playback studies of eight male frogs. Shown are the numbers of calls produced during NS, US and AUD periods (3 min each). The NS column represents control data before playback trials. Numbers in parentheses represent frog antiphonal responses that were precisely time-locked to the stimulus. Cells without values correspond to incomplete trials owing to the frog escaping from its calling site. **b**, Sound spectrogram (top left panel), waveform (bottom panel) and average amplitude spectrum (right panel) of a representative antiphonal response (produced 401 ms after the stimulus onset—shown between the two vertical cursors in the left panel), stimulated by the US components of the stimulus (390-ms long, to the left of the white vertical cursor). **c**, Ambient background noise recorded within 3 m of Tau Hua Creek in the absence of frog calls, recorded with a custom-made ultrasonic microphone having a high-pass cutoff frequency of 15 kHz with a roll-off of 10 dB per octave (hence the appearance of a progressive decrease in spectral energy from 0.1 to 15 kHz; see Methods). The noise has a nearly flat spectrum below 10 kHz when measured with a sound-level meter³.

¹Department of Molecular and Integrative Physiology & Beckman Institute, University of Illinois, Urbana, Illinois 61801, USA. ²Departments of Physiological Science and Ecology & Evolutionary Biology, University of California, Los Angeles, California 90095, USA. ³Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China. ⁴State Key Laboratory of Brain and Cognitive Science, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China.

*These authors contributed equally to this work.

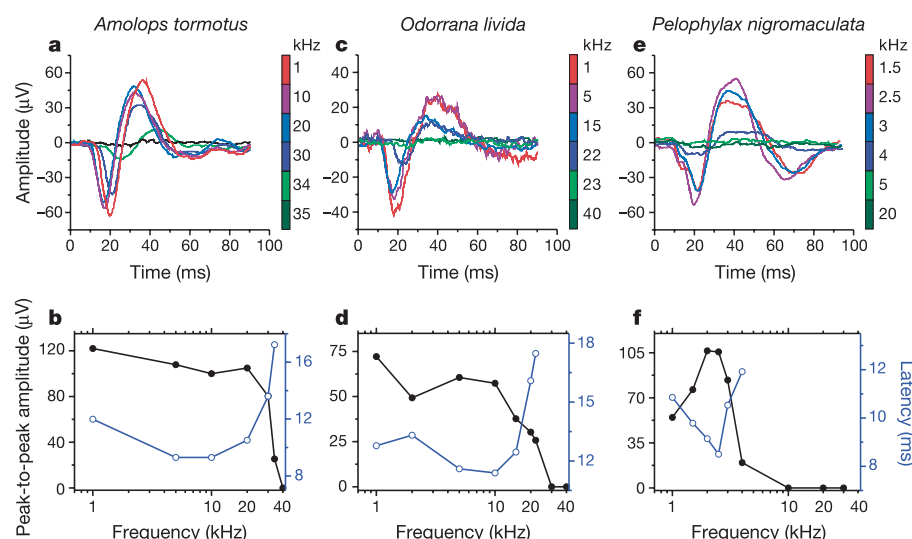


Figure 2 | Averaged auditory-evoked potential (AEP) data from the torus semicircularis validate the ultrasound sensitivity of *A. tormotus*. **a–f**, Shown are AEPs recorded from the torus semicircularis of *A. tormotus* (**a, b**), *O. livida* (**c, d**), and *P. nigromaculata* (**e, f**) in response to ten tone bursts over 1–40 kHz presented at a rate of 0.5 bursts s^{-1} .

a, c, e, Representative AEP waveforms. **b, d, f**, Corresponding peak-to-peak amplitudes ($N_1 - P_1$) of the AEPs (black) and N_1 latencies (blue) as a function of tone frequency. The N_1 latencies of the AEPs for the upper end of the species' hearing ranges were not measurable. N_1 and P_1 refer to the first negative and positive peaks of the AEP, respectively.

To validate the ultrasonic sensitivity of *A. tormotus* physiologically, we recorded auditory-evoked potentials (AEPs) from the torus semicircularis, the dominant midbrain auditory processing centre in the frog central nervous system. AEPs were consistently observed in response to tone bursts presented at 89 dB SPL from 1 to 34 kHz (Fig. 2a); no AEP was detectable for stimuli ≥ 35 kHz. The peak-to-peak AEP amplitudes were inversely correlated with latency (Fig. 2b). Because AEP latency decreases with increasing sound level above threshold, the shape of the latency curve approximates that of the species' audiogram.

We next isolated 30 single units from the torus semicircularis of 16 frogs and observed their tone-burst responses. Of these, 12 cells responded to tone bursts over a wide range of frequencies, including tones >20 kHz. For example, one tonic unit (Fig. 3a, right panels) fired maximally to a 20-kHz tone but also responded to tone bursts up to 27 kHz (Fig. 3a); a phasic-burst unit (Fig. 3b, right panels) responded best to a 10-kHz tone, with observable responses to tone bursts at 1–30 kHz (Fig. 3b). The distribution of the units' best frequencies is shown in Fig. 3c. Together, the AEP and single-unit data demonstrate the extraordinary ultrasonic sensitivity of *A. tormotus*.

The intense background noise from fast-flowing streams in Huangshan Hot Springs has a broad energy spectrum³, with a peak around 0.1 kHz, substantial amplitudes over 0.2–10 kHz, and a progressive decline from 11–22 kHz (Fig. 1c). Extending call frequencies into the ultrasonic range probably represents an adaptation to prevent the frog's vocal signals from being masked by the background noise⁴, similar to the high-frequency shift in the song of great tits living in urban areas⁶. Results of the present behavioural and electrophysiological studies support this hypothesis.

To determine whether sympatric frog species also show ultrasonic sensitivity, we recorded AEPs from the torus semicircularis of a large odorous frog (*Odorrana livida*) living in Huangshan Hot Springs. The AEP data from *O. livida* revealed that this frog also had the ability to detect ultrasound up to 22 kHz (Fig. 2c, d), but ultrasonic communication in this species remains to be shown.

In contrast, AEP recordings from the torus semicircularis of a black-spotted pond frog (*Pelophylax nigromaculata*), commonly found in rice fields and ponds throughout much of China⁷, revealed that its upper range of hearing was limited to 4–5 kHz (Fig. 2e, f). On the basis of these results and previously published studies^{1,2,8}, we suggest that ultrasonic hearing in frogs is: (1) probably limited to frog

species living in noisy environments, and (2) probably not due to artefacts in the acoustic stimulation system used in the physiological studies. Spectral analysis of the loudspeaker outputs revealed that the stimulus delivery system was quite linear; any subharmonics

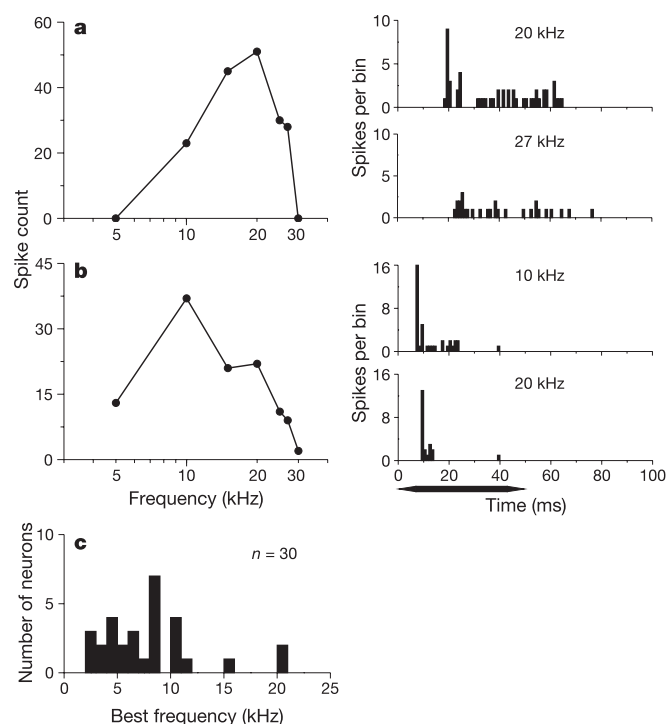


Figure 3 | Single-unit data from the torus semicircularis further confirm the ultrasound sensitivity of *A. tormotus*. **a**, Tone-burst responses from a tonic unit (see post-stimulus time histograms (PSTHs) in right panels) that responds to 10–27 kHz, with a best frequency of 20 kHz. **b**, Tone-burst responses from a phasic-burst unit (see PSTHs in right panels) that responds to 5–30 kHz, with a best frequency of 10 kHz. Vertical axes in **a, b** represent the total number of spikes in a 100-ms measurement window (with respect to the onset of tone bursts) at different frequencies. The horizontal bar below the four right panels represents the 'stimulus' starting at time zero. **c**, The distribution of best frequencies from single units.

generated by this system were at least 80 dB below the amplitude of the fundamental frequency of the stimulus.

In humans, ultrasound can be detected via bone conduction-mediated stimulation of the ear^{9–11} or auditory cortex^{12,13}. To determine whether the frog's ear is responsible for the ultrasonic sensitivity in *A. tormotus*, we carried out AEP recordings from the torus semicircularis of one frog under: (1) an 'intact' condition, with both ears unobstructed, or (2) an 'occluded' condition, with modelling clay covering the openings of both ear canals. The AEPs recorded under the intact condition (pre- and post-occlusion) were normal, with clear evidence of ultrasonic sensitivity (Fig. 4a), whereas ear occlusion abolished AEPs (Fig. 4b). Thus, ultrasonic sensitivity in *A. tormotus* is mediated by acoustic stimulation of the ear.

Frogs have two primary auditory organs: the amphibian papilla, which is sensitive to low and intermediate frequencies, and the basilar papilla, which is sensitive to high frequencies^{14,15}. In previous studies, the highest upper limit recorded for frog hearing as determined by the auditory sensitivity of nerve fibres innervating the basilar papilla is 8.2 kHz (ref. 16). As such, the particular organ in the frog's ear that contributes to ultrasonic sensitivity is unclear.

It has been suggested that the restricted hearing in frogs is largely attributed to the limited high-frequency response of their middle ear ossicles^{17–20}, owing to transmission loss resulting from flexion in the extracolumella–columella link¹⁹. Furthermore, with increasing frequency, the tympanic membrane vibration breaks up into higher vibration modes^{21,22}. Thus, for frogs to detect ultrasound they must circumvent these problems.

Males of *A. tormotus* have several highly unusual morphological features. First, their tympana are recessed (hence their common name) and invisible from the outside (Fig. 4c). Like mammals and unlike nearly all other frogs, they possess ear canals (Fig. 4c)—these have a resonant frequency of ~ 4.3 kHz (see Methods). Because the actual shape of the ear canal is complex, with distinct recesses, it might support secondary resonant frequencies, which may facilitate

high-frequency hearing. Second, recessed tympana shorten and therefore reduce the mass of the middle-ear ossicles. Third, the frog's tympanic membranes (Fig. 4c) are extremely thin (3–4 μm at the rim, 17–18 μm towards the centre, and 30–40 μm at the anchor point of the columellar footplate). Similarly, the eardrums of *O. livida* are thin and transparent, even though they are not embedded deep inside the ear cavities. Both low-mass ossicles and thin membranes facilitate transmission of high-frequency sounds to the inner ear.

At present, the specific contributions of the external, middle and inner ear (including the hair cells, the basilar papilla and the tectorial membrane) to the frog's auditory sensitivity, as well as the precise transmission characteristics of the external and middle ears, have yet to be determined. Elucidating these contributions represents fertile ground for research in comparative hearing and evolutionary biology. Additionally, although *A. tormotus* can detect and use ultrasound to communicate during male–male territorial interactions, it is unclear whether or not ultrasound is also involved in male–female interactions. Unlike the males, females of *A. tormotus* do not have recessed ears²³, and thus further research will be necessary to determine whether this ultrasonic hearing ability is sexually dimorphic.

METHODS

Acoustic playback experiments were carried out from 25 May to 2 June 2005, between 19:00 and 24:00 along the Tau Hua Creek in Huangshan Hot Springs, China. The WAV file of a pre-recorded call of *A. tormotus* (Fig. 1c in ref. 4) was stored on the flash memory of a custom-made playback unit, digitally filtered (low-pass or high-pass with a cutoff frequency of 20 kHz; slope 100 dB per octave), and then broadcast through an audible loudspeaker (Visaton DSM25FFL; pass band 2–22 kHz) or an ultrasonic loudspeaker (Polaroid; pass band 22–120 kHz) or both, at various playback levels. Ten stimuli were presented at a rate of 1 per 15 s over a 3-min period. The loudspeakers were mounted on a tripod and positioned 40–50 cm away from a calling frog.

The sound pressure levels (SPLs) of natural frog calls ranged from 65 to 92 dB SPL (fast reading r.m.s.) at a distance of 32 cm and an ambient noise level of 63 dB SPL (measured with a sound-level meter; GenRad 1982)³, depending on the call type and the centre frequency of the octave-band filter. At a distance of 40 cm, the r.m.s. SPL of the AUD components of the stimulus was 72.7 dB SPL, and that of the US components was 84.1 dB SPL; these were measured in the laboratory using a condenser microphone (Brüel and Kjaer 4135) and a precision measuring amplifier (Brüel and Kjaer 2610).

Frog movements were monitored visually under dim light, and their vocalizations were recorded using a PC-based digital recorder (PC-Tape) and a custom-made ultrasonic microphone having a flat frequency response from 15–120 kHz, with a roll-off of 10 dB per octave and 6 dB per octave at <15 kHz and >120 kHz, respectively²⁴; use of an ultrasonic microphone allowed detailed characterization of the spectral profile of the background noise and vocal signals in the ultrasonic range. Signals were digitized (16-bit A/D conversion) at a sampling rate of 256 kHz (8 $\times$ oversampling), saved as WAV files, analysed (fast Fourier transformed with 1024 points) and displayed using a custom-designed program⁴.

Electrophysiological experiments were carried out to determine the hearing range and sensitivity of the auditory system of several frog species: the concave-eared torrent frog (*A. tormotus*), the large odorous frog (*O. livida*) and the black-spotted pond frog (*P. nigromaculata*). Frogs were deeply anaesthetized by immersion in a 0.5% solution of tricaine methanesulphonate²⁵ and wrapped in cotton gauze. The skin on the dorsal surface of the head was incised, and a small hole made in the skull above the torus semicircularis. After surgery, the animal was placed inside a soundproof room and immobilized during the recording session with periodic injections of D-tubocurarine chloride (10 μg per g body weight).

Tone bursts (50–100-ms duration, 5-ms rise and fall times, presented at a rate of 0.5–1 pulse s^{-1}) were generated by a computer using A/D and D/A converters (Tucker Davis Technologies System 3), and broadcast from an ultrasonic loudspeaker (1–100 kHz; Tucker Davis Technologies ES-1) attached to a post positioned 10 cm from the frog's contralateral eardrum. The frequency response of the stimulation system was equalized to ± 6 dB over 2–40 kHz. SPLs were measured with a condenser microphone (Brüel and Kjaer 4135) and a sound-level meter (Brüel and Kjaer 2610).

Glass micropipette electrodes (tip diameter 1–2 μm) were used to record auditory-evoked potentials (AEPs) and single-unit activities from the torus

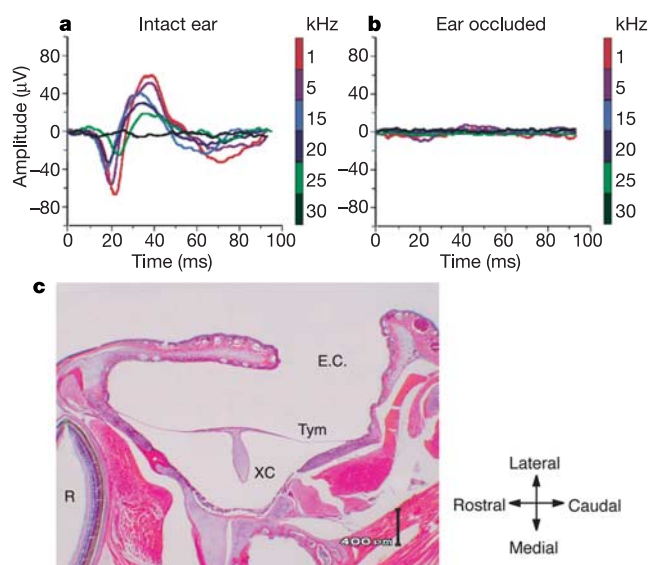


Figure 4 | The ear is responsible for ultrasound sensitivity in *A. tormotus*. **a, b**, AEP data recorded from the torus semicircularis of one *A. tormotus* in the intact condition (**a**) and occluded condition (**b**) show that ear occlusion abolishes the AEPs. **c**, Low-power photomicrograph ($\times 2$) of a horizontal section (stained with haematoxylin and eosin) through the right ear of a male of *A. tormotus*. See [http://www.beckman.uiuc.edu/profiles/feng/videos/for three-dimensional \(3D\) reconstruction of the ear canal](http://www.beckman.uiuc.edu/profiles/feng/videos/for three-dimensional (3D) reconstruction of the ear canal). Tym, tympanum; EC, ear cavity; XC, extracolumella. The opening of the ear cavity on the body surface is posterior to the retina (R); this opening was covered for the ear occlusion experiment described in **b**. Scale bar, 400 μm .

semicircularis in response to tone bursts at a constant and behaviourally relevant sound level of 80–90 dB SPL. Neural signals were amplified, monitored visually, extracted using BrainWare (Tucker Davis Technologies), stored on a hard drive and analysed off-line. AEPs were averaged over 10 trials; these measurements were repeated every 100 μ m along the dorsoventral extent of the torus semicircularis (3–5 electrode penetrations per frog). No attempt was made to determine the existence of tonotopy. Single-unit recordings involved a similar stimulation paradigm; each tone/intensity was presented 20 times to construct a post-stimulus time histogram (PSTH).

To evaluate the morphology of the ear canal of *A. tormotus*, three frogs were anaesthetized and decapitated. Heads were immersed in 10% formalin for 10 days, embedded in paraffin, sectioned (6- μ m slices) in the transverse, parasagittal and horizontal planes, and alternate slices were stained with luxol fast blue–cresyl violet and haematoxylin–eosin. Every fifth photographic image of the frog's transverse section was digitized. The digital files were stacked and aligned using the frog's sagittal profile as a guide to create a three-dimensional (3D) reconstruction of the ear canal (see video display at <http://www.beckman.uiuc.edu/profiles/feng/videofiles/>). The volume of the ear canal was calculated with Analyse AVW, an image-analysis software package. Assuming the ear canal is a simple Helmholtz resonator, its resonant frequency was calculated using the measured volume, and the area, radius and height of the opening (equations (8.1) and (8.16) in ref. 26).

Received 27 September; accepted 10 November 2005.

- Hauser, M. D. *The Evolution of Communication* 124–175 (MIT Press, Cambridge, 1996).
- Bradbury, J. W. & Vehrencamp, S. L. *Principles of Animal Communication* 161–173 (Sinauer Associates, Sunderland, 1998).
- Feng, A. S., Narins, P. M. & Xu, C. H. Vocal acrobatics in a Chinese frog, *Amolops tormotus*. *Naturwissenschaften* **89**, 352–356 (2002).
- Narins, P. M. et al. Old world frog and bird vocalizations contain prominent ultrasonic harmonics. *J. Acoust. Soc. Am.* **115**, 910–913 (2004).
- Sales, G. & Pye, D. *Ultrasonic Communication by Animals* 23–68 (Chapman and Hall, London, 1974).
- Slabbekoorn, H. & Peet, M. Birds sing at a higher pitch in urban noise. *Nature* **424**, 267 (2003).
- Fei, L. *Atlas of Amphibians of China* 162–163, 188–189, 230–231 (Henan Science and Technical Publisher, Henan, 1999).
- Rand, A. S. in *The Evolution of the Amphibian Auditory System* (eds Fritzsche, B., Ryan, M. J., Wilczynski, W., Hetherington, T. E. & Walkowiak, W.) 415–431 (John Wiley & Sons, New York, 1988).
- Pumphrey, R. J. Upper limit of frequency for human hearing. *Nature* **166**, 571 (1950).
- Corso, J. F. Bone-conduction thresholds for sonic and ultrasonic frequencies. *J. Acoust. Soc. Am.* **35**, 1738–1743 (1963).
- Dierooft, H. G. & Ertel, H. Some thoughts on the perception of ultrasonics by man. *Arch. Oto-Rhino-Laryn.* **209**, 277–290 (1975).
- Hosoi, H., Imaizumi, S., Sakaguchi, T., Tonoike, M. & Murata, K. Activation of the auditory cortex by ultrasound. *Lancet* **351**, 496–497 (1998).
- Imaizumi, S. et al. Ultrasound activates the auditory cortex of profoundly deaf subjects. *Neuroreport* **12**, 583–586 (2001).
- Feng, A. S., Narins, P. M. & Capranica, R. R. Three populations of primary auditory fibers in the bullfrog (*Rana catesbeiana*): their peripheral origins and frequency sensitivities. *J. Comp. Physiol.* **100**, 221–229 (1975).
- Lewis, E. R. & Narins, P. M. in *Comparative Hearing: Fish and Amphibians* (eds Fay, R. R. & Popper, A. N.) 218–268 (Springer, New York, 1999).
- Loftus-Hills, J. J. & Johnstone, B. M. Auditory function, communication, and the brain-evoked response in anuran amphibians. *J. Acoust. Soc. Am.* **47**, 1131–1138 (1970).
- Manley, G. A. The middle ear of the tokay gecko. *J. Comp. Physiol.* **81**, 239–250 (1972).
- Jørgensen, M. B. & Kannevorff, M. Middle ear transmission in the grass frog, *Rana temporaria*. *J. Comp. Physiol.* **182**, 59–64 (1998).
- Mason, M. J. & Narins, P. M. Vibrometric studies of the middle ear of the bullfrog *Rana catesbeiana* I. The extrastapes. *J. Exp. Biol.* **205**, 3153–3165 (2002).
- Mason, M. J. & Narins, P. M. Vibrometric studies of the middle ear of the bullfrog *Rana catesbeiana* II. The operculum. *J. Exp. Biol.* **205**, 3167–3176 (2002).
- Anson, M., Pinder, A. C., Keating, M. J. & Chung, S. H. Acoustic vibration of the amphibian eardrum studied by white noise analysis and holographic interferometry. *J. Acoust. Soc. Am.* **78**, 916–923 (1985).
- Purgue, A. P. Tympanic sound radiation in the bullfrog *Rana catesbeiana*. *J. Comp. Physiol.* **181**, 438–445 (1997).
- Chen, B.-H. *The Amphibian and Reptilian Fauna of Anhui* (Anhui Publishing House of Science and Technology, Hefei (Anhui), 1991).
- Siemers, B. M. & Schnitzler, H.-U. Echolocation signals reflect niche differentiation in five sympatric congeneric bat species. *Nature* **429**, 657–661 (2004).
- Ohr, E. A. Tricaine methanesulfonate-I: pH and its effects on anesthetic potency. *Comp. Biochem. Physiol. C* **54**, 13–17 (1976).
- Kinsler, L. E. & Frey, A. R. *Fundamentals of Acoustics* 186–200 (John Wiley, New York, 1962).

Acknowledgements We wish to thank X. Yu and J. Gao for their assistance in the field, E. Ko and B.-P. Umunna for analysis of the sound files, C. Park and J. Stelle for 3D reconstruction of the frog's ear, A. Schaub and B. Siemers for technical assistance with the recording and playback systems, L. Fei for identification of frog species, and M. Hauser, J. Simmons, T. Smith and R. Suthers for comments on the manuscript. This work was supported by grants from the National Institute on Deafness and Other Communication Disorders (to A.S.F. and P.M.N.), a grant from the National Science Foundation to A.S.F., a grant from the State Key Basic Research and Development Plan (China) to C.-H.X., and a grant from the National Natural Sciences Foundation (China) to J.-X.S.

Author Contributions A.S.F., P.M.N. and J.-X.S. were responsible for project planning. All authors (except Q.Q.) conducted the behavioural experiments, and A.S.F. and P.M.N. analysed the behavioural data. A.S.F., P.M.N., J.-X.S., Q.Q. and Z.-L.Y. conducted the electrophysiological experiments, and Q.Q. and Z.-L.Y. analysed the physiological data. A.S.F. performed the anatomical experiments and analysed the anatomical data.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.S.F. (afeng1@uiuc.edu).

Prioritizing global conservation efforts

Kerrie A. Wilson¹, Marissa F. McBride¹, Michael Bode¹ & Hugh P. Possingham¹

One of the most pressing issues facing the global conservation community is how to distribute limited resources between regions identified as priorities for biodiversity conservation^{1–3}. Approaches such as biodiversity hotspots⁴, endemic bird areas⁵ and ecoregions⁶ are used by international organizations to prioritize conservation efforts globally⁷. Although identifying priority regions is an important first step in solving this problem, it does not indicate how limited resources should be allocated between regions. Here we formulate how to allocate optimally conservation resources between regions identified as priorities for conservation—the ‘conservation resource allocation problem’. Stochastic dynamic programming is used to find the optimal schedule of resource allocation for small problems but is intractable for large problems owing to the ‘curse of dimensionality’⁸. We identify two easy-to-use and easy-to-interpret heuristics that closely approximate the optimal solution. We also show the importance of both correctly formulating the problem and using information on how investment returns change through time. Our conservation resource allocation approach can be applied at any spatial scale. We demonstrate the approach with an example of optimal resource allocation among five priority regions in Wallacea and Sundaland, the transition zone between Asia and Australasia.

Conservation organizations allocate resources to areas that have been identified as priorities for conservation investment^{3,7}. These priority regions are identified using information on relative biodiversity values, past or present threats to these values, and current levels of protection⁹. Species richness, or endemic species richness, is typically used to estimate the biodiversity value of a region¹⁰. The relative cost of conservation in different regions is ignored in the identification of priority regions despite evidence that its inclusion improves the cost-effectiveness of conservation prioritization^{11–15}.

Some international organizations rank these regions in terms of their priority for funding, but the approaches used to derive these rankings are not solutions of a properly formulated problem^{4,6,16}. If the objective is to maximize the total number of species conserved, then this objective is unlikely to be achieved if regions are prioritized only on the basis of species richness. This is because regions that are highly threatened but marginally less species-rich may lose many

species before being considered for conservation investment. Likewise, if the relative cost of investing in different regions is not taken into account, resources may be directed to expensive regions when the same amount of resources might have conserved more species if invested in regions with lower land-acquisition and management costs. The efficient allocation of conservation resources will be achieved only if the problem includes data on biodiversity, threat and cost, and is rigorously formulated.

Allocation of conservation resources, like any problem in decision theory, requires a broad goal, a specific objective, a set of constraints, a set of possible actions that form a strategy, and an understanding of the system dynamics provided by equations that link the actions and constraints to the objective¹ (see Methods). Here, the goal is to maximize biodiversity conservation through the creation of reserves, given ongoing habitat destruction and the constraint of a fixed budget. The best strategy—how much money to spend in each region each year—depends on endemic species richness, forest conversion rates (and the uncertainty associated with these rates), land cost and initial conditions (area of land currently reserved, converted or otherwise; see Table 1 and Supplementary Table). This decision theoretic formulation provides an explicit and transparent statement of the problem, which addresses the essential features of conservation resource allocation: biodiversity values, threats, costs, investment returns and data uncertainty.

We find an optimal resource allocation schedule using stochastic dynamic programming (SDP)^{8,17}. The SDP algorithm finds the optimal allocation decision each year given the current state of the system and possible events in the future¹⁷. Applying SDP to problems with more than a few regions is computationally intractable, so we investigate whether myopic heuristics (‘rules of thumb’ that look only one time-step ahead) can approximate the optimal solution. The two heuristics that we propose as approximations are ‘maximize short-term gain’ and ‘minimize short-term loss’ (see Methods). We compare the performance of these heuristics to priority setting approaches based on simple rankings, using a case study from Southeast Asia.

To illustrate our conservation resource allocation approach, we first compare the allocation of resources between two regions,

Table 1 | Biodiversity, threat and cost data for the five priority regions

Priority region	Area (km ²)	Forested area (km ²) in 1997	Reserved area (km ²) in 2003	No. of endemic bird species		Conversion rate (% yr ⁻¹)		Cost (US\$ km ⁻²)	
				Actual	Rank	Actual	Rank	Actual	Rank
Sumatra	475,746	164,303	84,901	18	4	2.3	2	95	2
Borneo	735,372	426,975	173,989	29	2	2.1	3	110	3
Sulawesi	187,530	79,509	68,150	67	1	2.4	1	76	1
Java/Bali	138,787	19,464	8,770	24	3	1.7	4	782	4
Southern peninsular Malaysia	131,598	58,500	29,221	4	5	1.2	5	2,746	5

¹The Ecology Centre, Schools of Integrative Biology and Physical Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia.

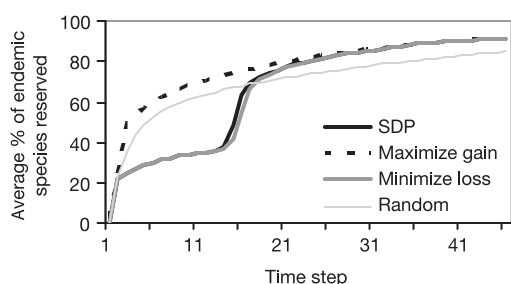


Figure 1 | Proportion of endemic species reserved through time in Borneo and Sumatra. The average proportion is calculated by four different resource allocation approaches: SDP, maximizing short-term gain, minimizing short-term loss, and random allocation between regions. The annual budget is US\$ 1 million and assumes no pre-existing reserves. The minimize short-term loss heuristic most closely approximates the optimal SDP solution; however, the maximize short-term gain heuristic results in the greatest number of endemic species reserved at early time steps.

Borneo and Sumatra, using the parameters in Table 1. We find that the optimal schedule is to allocate all resources to Sumatra for over a decade. Once all species occurring in Sumatra are conserved, investment is scheduled for Borneo. For this initial case study, the heuristic that minimizes short-term loss most closely approximates the optimal solution (Fig. 1). However, if there is uncertainty regarding our ability to invest in a region for the whole planning period, for example if funding ceases unexpectedly, then maximizing short-term gains is likely to result in the greatest number of species conserved. When we modify the problem to include a random probability of investment ceasing, the optimal allocation schedule more closely reflects the heuristic that maximizes short-term gain. The heuristic that maximizes short-term gain allocates funding to both regions simultaneously, in proportion to the marginal returns from investment (Fig. 2).

To explore the sensitivity of our results to the parameters, we assess each approach using different combinations of relative threat and relative endemic species richness for two hypothetical regions. Both heuristics perform well, but the heuristic that minimizes short-term loss is most similar to the optimal SDP solution and outperforms the other heuristic for most parameter sets (Fig. 3). The heuristic that maximizes short-term gain performs best when the threat levels of the two regions are similar, and performs poorly when the regions have very different threat levels but similar endemic species richness (Fig. 3a). The heuristic that minimizes short-term loss performs slightly worse than the optimal solution when both the relative level of threat and the endemic species richness of the two regions are very similar (Fig. 3b). If the annual budget is increased, land parcels are reserved at a faster rate. This mitigates forest conversion and, consequently, the difference between the approaches in the number of species conserved is reduced. When the relative cost of land

acquisition in each region is varied, the results are similar to those in Fig. 3 once the axes are adjusted to a species gain per dollar basis.

We next evaluate the performance of the heuristic algorithms, which we have shown to be close to optimal, for five priority regions from Southeast Asia. We compare the results of the algorithms with rankings based on endemic species richness, threat and cost (Table 1 and Fig. 4). This case study is used to illustrate our resource allocation approach, which can be applied to any number of priority regions for which a schedule for resource allocation is required. The approach can also be applied at any spatial scale, from global level problems to those at a local level. Our decision theory approach recommends initially investing all resources in Sulawesi and no other place until all the species occurring in Sulawesi are conserved. Only after this should investment proceed in Sumatra, Borneo and Java. After investment in Sulawesi ceases, the heuristic that maximizes short-term gain recommends roughly equal investment in Sumatra, Borneo and Java, and investment is scheduled last for Malaysia (Fig. 4a).

By contrast, ranking the five regions based on individual criteria does not provide an obvious schedule for resource allocation (Table 1). The three ranking criteria suggest that Sulawesi is the highest priority for conservation investment and Malaysia is the lowest. On the basis of only endemic bird richness, the rankings would recommend that investment should occur first in Sulawesi, second in Borneo, third in Java, fourth in Sumatra, and last in Malaysia. Although these simple rankings are not widely different from the results of our decision theory approach, there are discrepancies, which would occur more frequently for problems involving more regions. In addition, these rankings do not indicate the fraction of funds to allocate between the regions (nor whether funds should be allocated totally to a particular region or distributed between them). For example, the rankings could mean that investment should be directed towards Sulawesi until its species are conserved or, alternatively, that investments should be in proportion to the relative number of endemic species occurring in these regions. It is also not clear how these criteria should be combined to provide an allocation schedule that will maximize the protection of biodiversity. For example, if priority is determined only by endemic species richness, ignoring cost and threat, then Borneo's priority is overestimated. Similar confusion can arise if we prioritize only on threat or cost.

We have formulated the conservation resource allocation problem in a clear and transparent manner that involves defining an objective, identifying management actions, acknowledging constraints and incorporating uncertainty. Our problem formulation has five main simplifications. First, we have not accounted for the spatial variability within the priority regions and, consequently, the particular parcels where resources should be allocated are not identified. Second, our economic model is very simple. Third, we have not accounted for the temporally heterogeneous nature of land availability for reservation¹⁸. Fourth, we have used endemic bird species as a surrogate for biodiversity and assumed that numbers of endemic species reserved

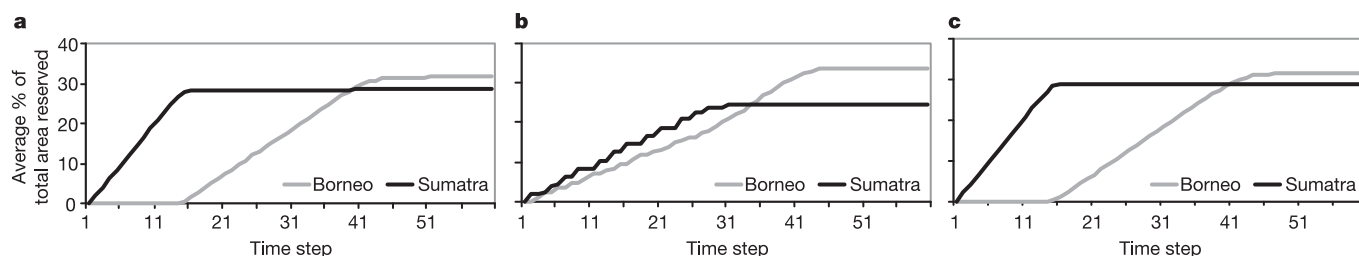


Figure 2 | Proportion of the total area of Borneo and Sumatra reserved through time. The average proportion of the total area reserved is calculated by three different resource allocation approaches. **a**, SDP. **b**, The heuristic that maximizes short-term gain. **c**, The heuristic that minimizes

short-term loss. The annual budget is US\$ 1 million and assumes no pre-existing reserves. The approach that maximizes short-term gain tends to allocate resources to both regions simultaneously, whereas the other, slightly superior, approaches allocate sequentially between regions.

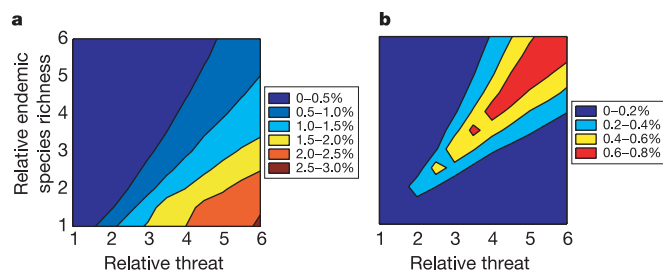


Figure 3 | Comparison of conservation performance of heuristic algorithms to the optimal solution. Two hypothetical regions are compared that differ in relative threat and number of endemic species. The performance of each heuristic is the percentage of endemic species not reserved relative to the optimum solution. **a**, Performance of the heuristic that maximizes short-term gain as compared with the optimal SDP solution. **b**, Performance of the heuristic that minimizes short-term loss as compared with the optimal SDP solution. The heuristic that minimizes short-term loss outperforms the heuristic that maximizes short-term gain for most parameter sets.

follows a species–area relationship. Last, we have assumed that the amount of resources invested in a region is directly proportional to the probability of species persisting. These assumptions can be relaxed within the framework that we present.

There is a need for computationally feasible and understandable algorithms that can deliver near-optimal solutions for large conservation resource allocation problems. The simple heuristics that we explore were developed to solve a properly formulated problem, perform surprisingly well relative to the optimal SDP solution, and are superior to simple ranking approaches. Minimizing short-term loss most closely approximates the optimal allocation schedule and maximizing short-term gain is close to optimal despite ignoring threat, although it underperforms if threat levels are very different. Under extreme uncertainty, maximizing short-term gain is the most risk-averse approach as it provides a buffer against an uncertain investment future.

We recommend that conservation organizations maximize short-term gain, unless the regions of concern have similar endemic species richness and very different levels of threat. In such circumstances,

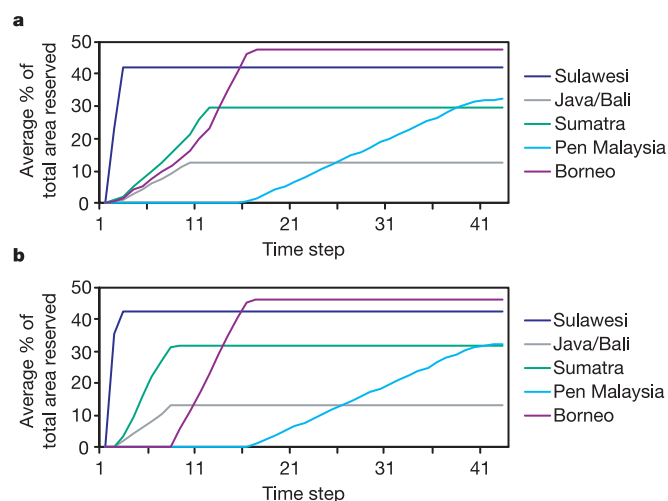


Figure 4 | Proportion of five priority regions from Southeast Asia reserved through time. The average proportion of the total area reserved is determined using two resource allocation heuristics. **a**, Heuristic that maximizes short-term gain. **b**, Heuristic that minimizes short-term loss. The annual budget is US\$1 million and assumes no pre-existing reserves. In both cases resources are initially allocated completely to Sulawesi and investment is directed to southern peninsular Malaysia only after investment has occurred in Sumatra, Java/Bali and Borneo.

resource allocation should minimize short-term loss. We argue that conservation investments should be evaluated as any investment is evaluated: that is, with a clearly defined objective and an assessment of how well the returns from the investment meet this objective. Responsible conservation organizations and international agencies should consider embracing a decision theoretic approach when scheduling the allocation of conservation resources.

METHODS

The first step in formulating the conservation resource allocation problem is to define a quantifiable objective. Our objective is to maximize the number of endemic species remaining across all regions when habitat conversion ceases because there is no unreserved or unconverted land (see Supplementary Methods ‘Problem formulation’). We assume that, for each region, the number of endemic species conserved per unit area is a monotonically decreasing function of the area reserved. Therefore, our conservation returns in each region diminish with increasing investment. We model this relationship using a species–area curve^{19–21} (see Supplementary Methods (ii)). In principle, any relationship between area and endemic species conserved could be used. The next step in formulating the problem is determining what actions are possible, in this case what fraction of the budget to allocate to each region each year. Our decisions are thus constrained by a fixed annual budget. Although we recognize that funds for conservation can be directed towards many kinds of activity (for example, restoration programs, the purchase of forestry concessions and species recovery programs), we focus on the acquisition of land for reservation. We assume that each land parcel can be classified as reserved, available for reservation or converted (anthropogenically altered and assumed no longer suitable habitat for the endemic species of the region). Threat is modelled by assuming that a constant proportion of available parcels in each region are converted each year. To incorporate the uncertainty associated with parcel loss^{22–24} (see Supplementary Methods (iii)), conversion is represented as a stochastic process with a binomial distribution. We estimate the cost of reservation in each region using statistical models^{12,25} (see Supplementary Methods (iv)).

We use SDP and two myopic heuristics to determine how many parcels to reserve in each region each year. We compare these results to a random acquisition process. Once the optimal solution is obtained, we forward-simulate the resulting acquisition schedule 10,000 times for each parameter set to calculate the expected number of species conserved. The solutions found using SDP are optimal in the face of uncertainty^{18,26–28}. Owing to an exponentially increasing state space, however, ‘the curse of dimensionality’ limits SDP to problems with few regions. The maximize short-term gain heuristic selects parcels for reservation that result in the greatest increase in the number of endemic species conserved. This heuristic ignores threat and is myopic, considering only the short-term future when selecting the next parcel to reserve and not all possible futures as the SDP algorithm does. The minimize short-term loss heuristic is also myopic: it selects parcels that will minimize the expected loss of species from the system in the next time step (see Supplementary Methods (i)).

Received 27 September; accepted 24 October 2005.

- Possingham, H. P., Andelman, S. J., Noon, B. R., Trombulak, S. & Pulliam, H. R. in *Research Priorities for Conservation Biology* (eds Orians, G. & Soulé, M.) 225–244 (Island Press, Washington DC, 2001).
- Pimm, S. L. *et al.* Can we defy nature's end? *Science* **293**, 2207–2208 (2001).
- Dalton, R. Biodiversity cash aimed at hotspots. *Nature* **406**, 818 (2000).
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
- Stattersfield, A. J., Crosby, M. J., Long, A. J. & Wege, D. C. *Endemic Bird Areas of the World: Priorities for Biodiversity Conservation* (Birdlife International, Cambridge, UK, 1998).
- Olson, D. M. & Dinerstein, E. The global 200: priority ecoregions for global conservation. *Ann. MO Bot. Gard.* **89**, 199–224 (2002).
- Halpern, B. S. *et al.* Gaps and mismatches between global conservation priorities and spending. *Conserv. Biol.* (in the press).
- Bellman, R. *Dynamic Programming* (Princeton Univ. Press, Princeton, NJ, 1957).
- Wilson, K. A. *et al.* Measuring and incorporating vulnerability into conservation planning. *Environ. Manag.* **35**, 527–543 (2005).
- Orme, C. D. L. *et al.* Global hotspots of species richness are not congruent with endemism or threat. *Nature* **436**, 1016–1019 (2005).
- Balmford, A., Gaston, K. J., Rodrigues, A. S. L. & James, A. Integrating costs of conservation into international priority setting. *Conserv. Biol.* **14**, 1–9 (2000).
- Moore, J., Balmford, A., Allnutt, T. & Burgess, N. Integrating costs into conservation planning across Africa. *Biol. Conserv.* **117**, 343–350 (2004).
- Frazee, S. R., Cowling, R. M., Pressey, R. L., Turpie, J. K. & Lindenberger, N. Estimating the costs of conserving a biodiversity hotspot: a case-study of the Cape Floristic Region, South Africa. *Biol. Conserv.* **112**, 275–290 (2003).

14. Possingham, H. & Wilson, K. Biodiversity—turning up the heat on hotspots. *Nature* **436**, 919–920 (2005).
15. Odling-Smee, L. Conservation: dollars and sense. *Nature* **437**, 614–616 (2005).
16. Groves, C. *et al.* *Designing a Geography of Hope. A Practitioner's Guide for Ecoregional Conservation Planning* (The Nature Conservancy, Washington DC, 2000).
17. Clark, C. W. & Mangel, M. *Dynamic State Variable Models in Ecology* (Oxford Univ. Press, New York, 2000).
18. Meir, E., Andelman, S. & Possingham, H. P. Does conservation planning matter in a dynamic and uncertain world? *Ecol. Lett.* **7**, 615–622 (2004).
19. Rosenzweig, M. *Species Diversity in Space and Time* (Cambridge Univ. Press, Cambridge, UK, 1995).
20. Brooks, T. M. *et al.* Habitat loss and extinction in the hotspots of biodiversity. *Conserv. Biol.* **16**, 909–923 (2002).
21. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton Univ. Press, Princeton, NJ, 1967).
22. Achard, F. *et al.* Determination of deforestation rates of the world's humid tropical forests. *Science* **297**, 999–1002 (2002).
23. Linkie, M., Smith, R. J. & Leader-Williams, N. Mapping and predicting deforestation patterns in the lowlands of Sumatra. *Biodivers. Conserv.* **13**, 1809–1818 (2004).
24. Curran, L. M. *et al.* Lowland forest loss in protected areas in Indonesian Borneo. *Science* **303**, 1000–1003 (2004).
25. Balmford, A., Gaston, K. J., Blyth, S., James, A. & Kapos, V. Global variation in terrestrial conservation costs, conservation benefits, and unmet conservation needs. *Proc. Natl Acad. Sci. USA* **100**, 1046–1050 (2003).
26. Possingham, H., Day, J., Goldfinch, M. & Salzborn, F. in *Decision Sciences: Tools for Today, 12th Australian Operations Research Conference* (eds Sutton, D., Cousins, E. & Pearce, C.) 536–545 (ASOR, Univ. Adelaide, Adelaide, SA, 1993).
27. Costello, C. & Polasky, S. Dynamic reserve site selection. *Res. Energy Econ.* **26**, 157–174 (2004).
28. Drechsler, M. Probabilistic approaches to scheduling reserve selection. *Biol. Conserv.* **122**, 253–262 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Elkin, T. Martin and E. Game for comments on the manuscript; and P. Kareiva, S. Polasky, R. L. Pressey, S. Andelman and B. Murdoch for discussions. The work was supported by The University of Queensland and grants from the Australian Research Council (to H.P.P., M. A. McCarthy and R. L. Pressey).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.P.P. (h.possingham@uq.edu.au) or K.A.W. (k.wilson2@uq.edu.au).

Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts

Devin R. Scannell^{1*}, Kevin P. Byrne^{1*}, Jonathan L. Gordon¹, Simon Wong¹ & Kenneth H. Wolfe¹

A whole-genome duplication occurred in a shared ancestor of the yeast species *Saccharomyces cerevisiae*, *Saccharomyces castellii* and *Candida glabrata*. Here we trace the subsequent losses of duplicated genes, and show that the pattern of loss differs among the three species at 20% of all loci. For example, several transcription factor genes, including *STE12*, *TEC1*, *TUP1* and *MCM1*, are single-copy in *S. cerevisiae* but are retained in duplicate in *S. castellii* and *C. glabrata*. At many loci, different species have lost different members of a duplicated gene pair, so that 4–7% of single-copy genes compared between any two species are not orthologues. This pattern of gene loss provides strong evidence for speciation through a version of the Bateson–Dobzhansky–Muller mechanism, in which the loss of alternative copies of duplicated genes leads to reproductive isolation^{1,2}. We show that the lineages leading to the three species diverged shortly after the whole-genome duplication, during a period of precipitous gene loss. The set of loci at which single-copy paralogues are retained is biased towards genes involved in ribosome biogenesis and genes that evolve slowly, consistent with the hypothesis that reciprocal gene loss is more likely to occur between duplicated genes that are functionally indistinguishable. We propose a simple, unified model in which a single mechanism—passive gene loss—enabled whole-genome duplication and led to the rapid emergence of new yeast species.

We used the Yeast Gene Order Browser (YGOB, ref 3) to compare six yeast species, three of which diverged after their common ancestor experienced a whole-genome duplication (WGD), and three of

which diverged from this lineage before the WGD. The YGOB compares pairs of genomic regions from post-WGD species (*S. cerevisiae*^{4,5}, *S. castellii*⁶ and *C. glabrata*⁷) to single genomic regions in pre-WGD species (*Kluyveromyces waltii*⁸, *Kluyveromyces lactis*⁷ and *Ashbya gossypii*⁹) (Fig. 1). We use the term ‘ancestral locus’ to describe a locus in a pre-WGD species, or the corresponding duplicated pair of loci in a post-WGD species (that is, a column in Fig. 1). Synteny conservation enabled us to determine unambiguously whether each of 2,723 ancestral loci was retained in one or two copies in each post-WGD genome. If only one copy remained, the syntenic context allowed orthologues to be distinguished from paralogues (Fig. 1).

The fate of an ancestral locus among the three post-WGD species can be classified into one of 14 possible patterns (Fig. 2). The most common pattern (Class 4, seen at 1,957 ancestral loci or 72% of the total) is that all three species have lost the same (orthologous) copy of the gene, such as in the *LYS2* column in Fig. 1. For clarity, we show this as three separate losses in Fig. 2, but a loss could have occurred in the ancestor of two or three of the species. Of the ancestral loci, 210 (8%) remain duplicated in all three post-WGD species (Class 0). The other 556 ancestral loci (20%) have had variable fates among the three post-WGD species, which indicates that the consequences of WGD were still being sorted out when these lineages diverged. A striking example is the set of 18 genes that are single-copy in *S. cerevisiae* but double-copy in both *S. castellii* and *C. glabrata* (Class 1B). Transcription factors are disproportionately over-represented in this group (it includes *STE12*, *TUP1*, *GAL11*, *GCR2*,

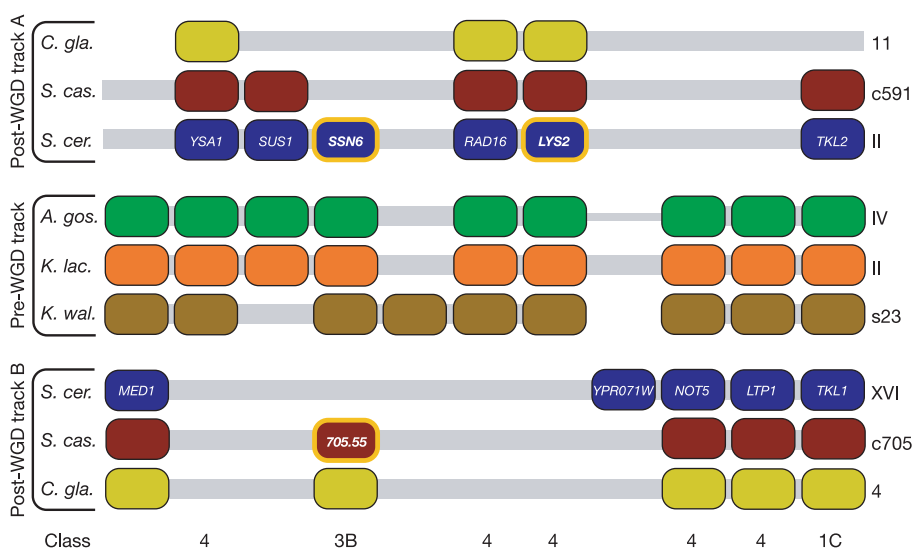


Figure 1 | Gene order relationships in the region around *S. cerevisiae* *SSN6* and its homologues. Relationships based on YGOB output (<http://wolfe.gen.tcd.ie/ygob>). Coloured boxes represent genes and are not drawn to scale. Chromosomal regions from each pre-WGD species are represented by one horizontal track each. The two corresponding regions in each post-WGD species are represented by two tracks (A and B) at the top and bottom. Chromosome, contig or scaffold numbers are indicated on the right. Homologous genes are arranged in columns. Thick grey horizontal bars connect genes that are immediate neighbours in the genome. Codes below columns indicate the gene loss class for that ancestral locus, as used in Fig. 2. Columns without codes did not meet the criteria for scoring.

¹Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland.

*These authors contributed equally to this work.

SFP1, *YAP3* and *TYE7*; $P = 0.001$ by Fisher's test), which suggests that the transcriptional regulatory network in *S. cerevisiae* is simpler than in the other yeasts (Supplementary Table S1). *MCMI* and *TEC1* are also in a 1:2:2 relationship among the post-WGD genomes, but these two loci were not counted in Fig. 2 because the syntenic context around them is not completely conserved.

S. cerevisiae *SSN6* (also known as *CYC8*) and *S. castellii* gene *705.55* are an example of single-copy paralogues (Fig. 1). This situation arises when opposite members of a gene pair are lost in two daughter species. Between *S. cerevisiae* and *S. castellii*, 176 of the 2,723 loci we surveyed (6.4%; Classes 2E, 3A and 3B in Fig. 2) show this pattern of reciprocal gene loss. Reciprocal gene loss is a particular form of reciprocal silencing¹ or divergent resolution^{2,10} of duplicated genes, and is a property of a pair of genomes. Similarly, there are 198 reciprocal gene-loss loci between *C. glabrata* and *S. castellii* (7.3%), and 100 between *S. cerevisiae* and *C. glabrata* (3.7%). Thus, a significant minority of genes that are mutual best BLASTP hits between the post-WGD genomes are not orthologues. More importantly, the process of reciprocal gene loss has the effect of changing the location of the functional copy of a gene^{1,2}. For instance, *S. castellii* effectively carries a null allele at its locus orthologous to *SSN6*, and *S. cerevisiae* has a null allele orthologous to gene *705.55* (Fig. 1). If this were the only difference between these two species and they formed a hybrid, the hybrid would be likely to have low fitness because one-quarter of its spores would lack a functional copy of both *SSN6* and gene *705.55* (*S. cerevisiae* *ssn6* mutants are defective in respiratory growth and sporulation). In fact, 66 of the 176 loci that have undergone reciprocal gene loss between *S. cerevisiae* and *S. castellii* involve essential *S. cerevisiae* genes, so the spore viability of the hypothetical hybrid is reduced to approximately $(0.75)^{66}$ or 6×10^{-9} as a consequence of essential genes alone. Viability will be reduced further by reciprocal gene loss at loci that were not scored in Fig. 2 owing to inadequate synteny conservation (about half the genome), and at loci such as *SSN6* that are not essential but still contribute to fitness. The number of reciprocal losses observed among the post-WGD species is ample to account for their reproductive isolation,

notwithstanding the contributions of mechanisms such as inter-chromosomal rearrangement^{11,12} and mismatch repair^{13,14}.

The situation described above for the *SSN6* and *705.55* genes is a special case of Bateson–Dobzhansky–Muller (BDM) interspecific genomic incompatibility¹⁵. The BDM model proposes that negative epistatic interactions between two loci can reduce the fitness of a hybrid. Werth and Windham¹ and Lynch and Force² applied the BDM model to duplicated genes, hypothesizing that reciprocal loss (or silencing) of different copies in two species would create a BDM incompatibility, leading to reduced hybrid fitness. Reciprocal gene loss at multiple loci could lead to reproductive isolation, and where many duplicated genes exist (as in a polyploid) there is the potential for successive nested speciation events to occur^{1,2,10}.

To investigate whether reciprocal gene loss was involved in the establishment of reproductive isolation among the post-WGD lineages, we determined the timing of gene losses by estimating the number of duplicated genes surviving at each node on the lineage leading to *S. cerevisiae* (Fig. 3). To increase the resolution of this analysis we included data from *S. bayanus*¹⁶, a close relative of *S. cerevisiae*. (Reproductive isolation between these two species is due to processes other than reciprocal gene loss¹².) We expressed the ages of the nodes as a proportion of the time (T) since the initial divergence of gene pairs created by WGD (see Supplementary Notes S2 and S3). We then estimated the numbers of genes still duplicated in the common ancestors of *S. cerevisiae* and each of *S. bayanus*, *C. glabrata* and *S. castellii* using two methods: parsimony (which gives the minimum number of genes that must have been retained in duplicate) and a model-based approach (Supplementary Note S4). We consider the latter to be more realistic because it allows for parallel gene losses in different lineages.

The parsimony and model-based methods both show a precipitous loss of duplicated genes in the time interval between WGD and the first speciation event (Fig. 3b, c). Both methods also show that the fraction of genes retained in duplicate declined appreciably (from 47% to 32% according to the model-based method) in the interval between the first (*S. castellii*) and the second (*C. glabrata*) speciation, even though this corresponds to a very short time period. From this we conclude that gene loss was still occurring rapidly during the emergence of the post-WGD lineages. Moreover, because reciprocal gene loss (by definition) cannot have occurred before *S. castellii* diverged from the other post-WGD lineages, and the number of gene losses on the right-hand side of the curve is very few (*S. bayanus* differs from *S. cerevisiae* at only two of the scored ancestral loci), the vast majority of reciprocal losses must have occurred at around the time of the two speciation events. In fact, we estimate that two-thirds of all reciprocal gene-loss events occurred between the time of *S. castellii* divergence and time $0.337T$ (Fig. 3b). The reproductive barriers imposed on these species by reciprocal gene loss are therefore not recent reinforcements but were erected contemporaneously with speciation.

The fate awaiting most gene pairs formed by WGD was that the duplication was resolved by deleting one gene copy (Fig. 2). If the two copies were functionally identical, we would expect the 'choice' of which copy to delete to be arbitrary. This hypothesis can be tested at ancestral loci that have been resolved independently in more than one post-WGD lineage. We find that in cases of two independent losses, the two retained genes are more often orthologues than paralogues (in Fig. 2, compare Class 2D to 2C, and 2F to 2E; χ^2 -test of homogeneity, $P < 0.05$ for each). A possible explanation for the excess of convergent losses is that at some loci the two copies were not functionally identical, and that the same (better-functioning) copy was retained on both occasions. In contrast, the fact that divergent resolution is seen at some other loci suggests that the choice of survivor at those loci was arbitrary (Classes 2A, 2C, 2E and 3). These observations can be reconciled if some pairs of genes were functionally indistinguishable at the time the duplication was resolved (in which case either copy could be retained) but others were

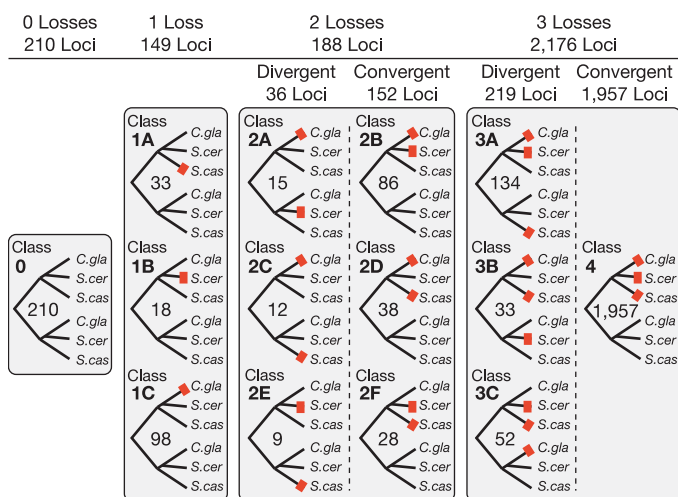


Figure 2 | Classes of gene loss pattern among 2,723 ancestral loci in *S. cerevisiae*, *S. castellii* and *C. glabrata*, and their frequencies. Red marks denote gene absence and are used to group ancestral loci into 14 gene-loss classes, described by schematic trees showing the fates of orthologous and paralogous genes. The number of ancestral loci in each gene-loss class is shown in the centre of the tree. The two sets of species names in each tree denote tracks A and B in arbitrary order. In some cases, the absence of a gene copy in two or more species may be due to a single gene-loss event on a shared branch, but this does not affect classification. Convergent classes are those in which all genes lost are orthologues; divergent classes involve some loss of paralogues in different species.

functionally distinct (so that a particular copy was preferred by selection).

Differences in the performance of a function can only have been due to sequence differences between the gene copies themselves or between their *cis*-regulatory regions. This sequence divergence must have accumulated in the time between WGD and gene loss, or, if the WGD was an allopolyploidy, have been inherited from parental species. Therefore, neutral gene loss (which results in divergent resolution half of the time) is expected to be more frequent at ancestral loci that are slowly evolving or involved in highly conserved biological processes where the potential for functional divergence is low. We tested this prediction and indeed find that loci in Class 3 (all of which underwent reciprocal gene loss between two species) evolve on average 30% slower than Class 4 (for which no reciprocal gene loss occurred) (Supplementary Note S5; Wilcoxon rank-sum test $P < 10^{-14}$). Moreover, Gene Ontology terms such as 'ribosomal RNA processing', 'ribosome biogenesis' and 'RNA binding' are disproportionately over-represented among Class 3 loci, as are proteins that are localized to the nucleolus¹⁷ and proteins in complexes that bind RNAs¹⁸ (Supplementary Table S1 and Note S5). Finally, we also find that genes for small nucleolar (sno)RNAs, many of which function in ribosomal RNA processing, have undergone reciprocal gene loss unusually frequently (Supplementary Note S5). Thus, the set of reciprocal gene-loss loci seems biased towards those with functions most likely to be conserved between duplicates. This functional bias increases the potential contribution of reciprocal gene-loss loci to reproductive isolation, because 40% of the Class 3 loci are essential¹⁹ in *S. cerevisiae*, compared with 20% of Class 4 loci ($P < 10^{-10}$, χ^2 -test).

The passive loss of genes from genomes in which there is no selection to retain them is a familiar phenomenon in molecular evolution^{20,21}. We suggest that passive gene loss is the likely mechanism of the original WGD event in yeast. Our model (Fig. 4) begins with two haploid cells fusing to form a diploid. If the haploids are from different species, or differ by a chromosomal rearrangement, or carry particular mutations, the resulting diploid may be unable to form viable spores but still able to divide mitotically. If the diploid cell lineage continues to divide mitotically for many generations, it can start to lose one allele from every locus that is not haploinsufficient. During this process, there is nothing to prevent an allele at the *MAT* locus from being deleted, in which case the cell will behave as a haploid. It can switch mating type, undergo mother–daughter mating, form a diploid, and so regain fertility²². Former alleles become separate loci, each of which is homozygous. Continuing loss of redundant gene copies will result in separate lineages that are self-fertile but reproductively isolated from one another by reciprocal gene loss (Fig. 4).

Our results provide evidence that reciprocal gene loss at multiple ancestrally duplicated genes may lead to speciation, as has previously been hypothesized (but not demonstrated) for polyploid plants^{1,23} and fish^{10,24}. Indeed, because we have shown that reciprocal gene loss is implicated in the emergence of three different lineages, our data support the feature of the modified BDM mechanism^{1,2} that most distinguishes it from other theories of reproductive isolation: the ease with which it accounts for multiple speciation events. Finally, by showing that slowly evolving genes and those involved in very fundamental processes are the ones most likely to undergo reciprocal

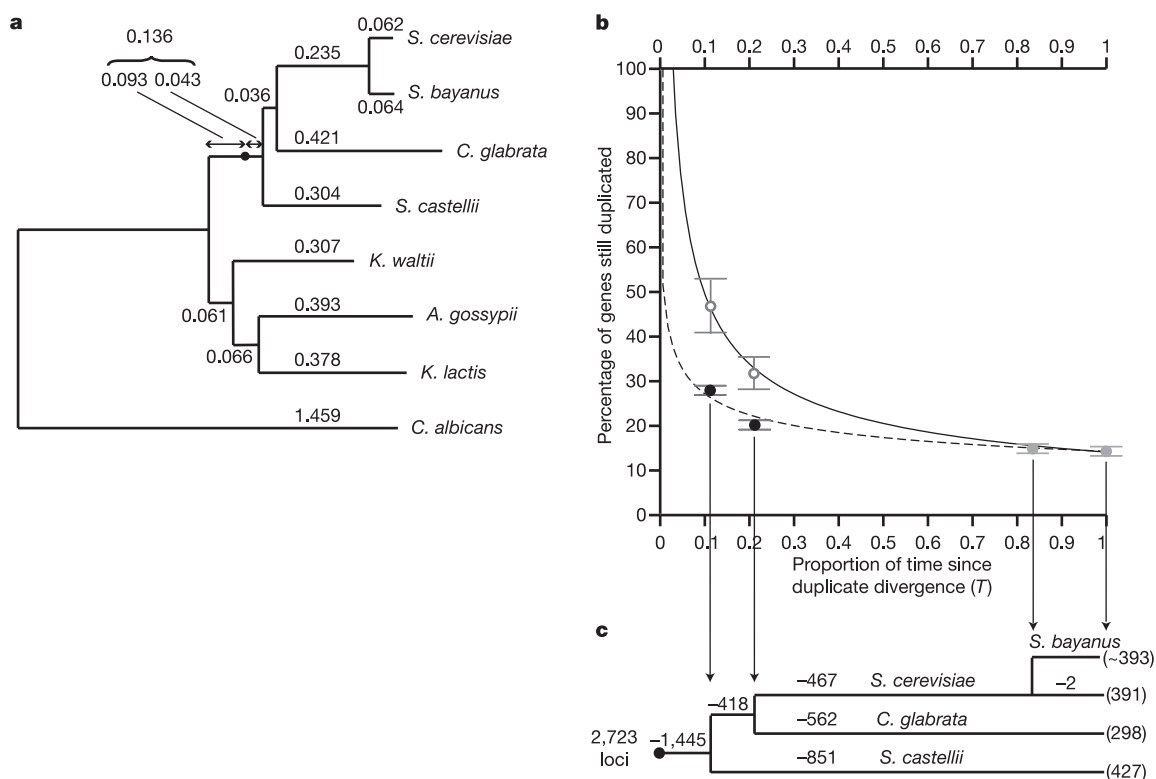


Figure 3 | Time course of duplicated gene loss following WGD. a, Tree reconstructed from 909 protein sequences using a constrained topology (Supplementary Note S2) and branch-length estimation by maximum likelihood. The black dot indicates the initial divergence of duplicates created by WGD (Supplementary Note S3). **b**, Gene-loss curves estimated by the model-based method (open circles and solid curve; Supplementary Note S4) and by parsimony (black circles and dashed curve). Filled grey circles are common to both methods and show the percentages of loci duplicated in *S. cerevisiae* and its common ancestor with *S. bayanus*. The horizontal scale

represents the time from the initial divergence of duplicates created by WGD (0T) to the present (1T) and is derived from the tree in **a** assuming a molecular clock (Supplementary Note S3). Power-law curves were fitted to the data²⁵. Standard errors for x (all $< 2\%$; omitted for clarity) and y values were estimated by bootstrapping. **c**, Numbers of genes lost on each branch leading to post-WGD species, as inferred by the model-based method. The current numbers of duplicates remaining in each post-WGD genome are shown in parentheses. All numbers refer to the 2,723 loci summarized in Fig. 2.

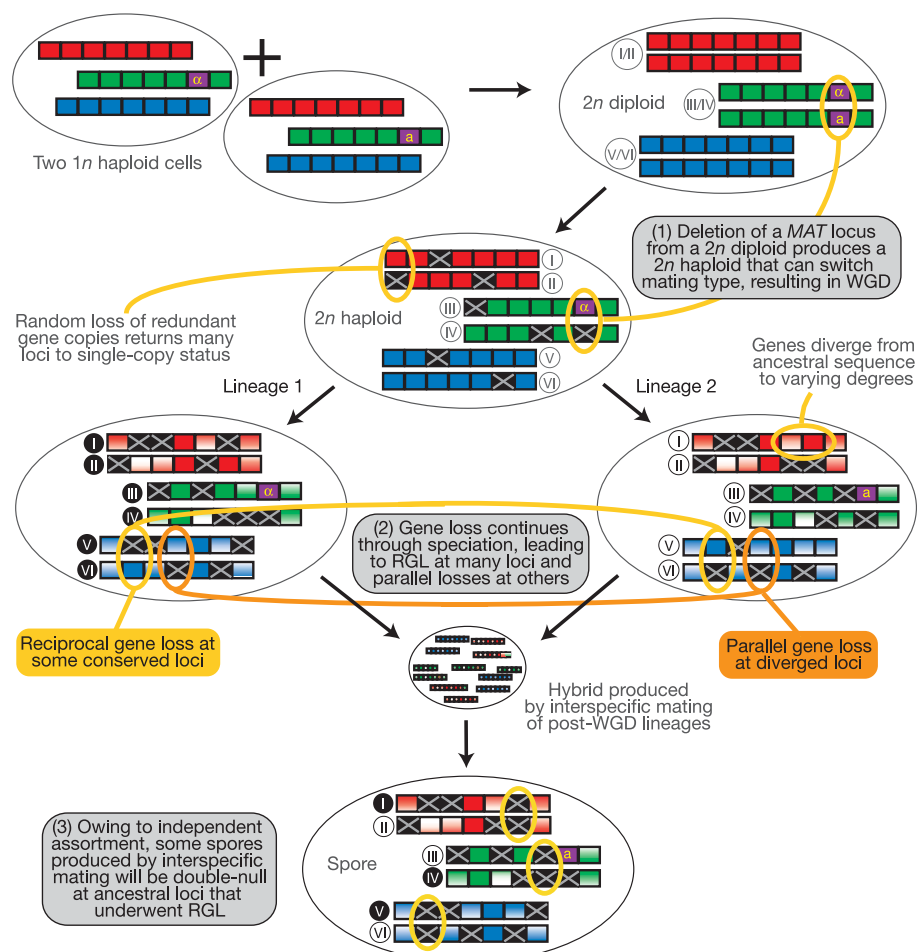


Figure 4 | Model of passive gene loss as a mechanism for WGD and establishment of reproductively isolated lineages. Individual steps are discussed in the main text. Ovals represent yeast cells. Genes are shown as red, green or blue boxes (except for the *MAT* locus, shown in purple), and are

arranged horizontally as chromosomes. Grey 'X' symbols indicate genes that have been deleted. Roman numbering of chromosomes is used to indicate the parent of origin where relevant. Features relevant to each step are circled in yellow or orange. RGL, reciprocal gene loss.

gene loss, our study leads to the conclusion that these genes, which individually are among the most conservative in the genome, may collectively be responsible for the most radical of evolutionary events.

METHODS

We used the YGOB engine³ to assess the status and syntenic conservation of loci in *S. cerevisiae*, *S. castellii* and *C. glabrata*. Each ancestral locus (that is, a column in Fig. 1, corresponding to two genomic sites in post-WGD species and one site in pre-WGD species) was scored up to 18 times: on tracks A and B in each of the three post-WGD species, and comparing against each of the three pre-WGD genomes. On the basis of homology and syntenic context, the status of each of the six genomic sites in the post-WGD species was designated as one of (1) gene unambiguously present, (2) gene unambiguously absent, (3) gene present but with insufficient syntenic support, or (4) gene absent but with insufficient syntenic support. Loci were retained for further analysis if presence or absence could be determined unambiguously on both tracks in all three post-WGD species and if the scoring against all three pre-WGD genomes was not contradictory. This yielded reliable information for 2,723 ancestral loci, as summarized in Fig. 2. The scoring protocol and our implementation are described in ref. 3. We ignored a small number of ancestral loci for which one of the post-WGD species retained neither gene copy. *S. bayanus* was scored relative to the 2,723 ancestral loci in *S. cerevisiae* because their genomes are almost completely co-linear. In *S. bayanus*, 2,631 loci had conserved syntenic context (by the criteria above), and manual inspection of candidates generated by the YGOB engine revealed just two differences between *S. bayanus* and *S. cerevisiae* (*S. bayanus* has retained paralogues as well as orthologues of *HEK2* and *YAT1*).

Received 5 August; accepted 22 December 2005.

1. Werth, C. R. & Windham, M. D. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *Am. Nat.* **137**, 515–526 (1991).
2. Lynch, M. & Force, A. G. The origin of interspecies genomic incompatibility via gene duplication. *Am. Nat.* **156**, 590–605 (2000).
3. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15**, 1456–1461 (2005).
4. Goffeau, A. *et al.* The Yeast Genome Directory. *Nature* **387** (Suppl.), 5–105 (1997).
5. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
6. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
7. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430**, 35–44 (2004).
8. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–624 (2004).
9. Dietrich, F. S. *et al.* The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304**, 304–307 (2004).
10. Taylor, J. S., Van de Peer, Y. & Meyer, A. Genome duplication, divergent resolution and speciation. *Trends Genet.* **17**, 299–301 (2001).
11. Fischer, G., Neuvéglise, C., Durrrens, P., Gaillardin, C. & Dujon, B. Evolution of gene order in the genomes of two related yeast species. *Genome Res.* **11**, 2009–2019 (2001).
12. Delneri, D. *et al.* Engineering evolution to study speciation in yeasts. *Nature* **422**, 68–72 (2003).
13. Greig, D., Louis, E. J., Borts, R. H. & Travisano, M. Hybrid speciation in experimental populations of yeast. *Science* **298**, 1773–1775 (2002).
14. Hunter, N., Chambers, S. R., Louis, E. J. & Borts, R. H. The mismatch repair

- system contributes to meiotic sterility in an interspecific yeast hybrid. *EMBO J.* **15**, 1726–1733 (1996).
15. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer, Sunderland, Massachusetts, 2004).
 16. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
 17. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
 18. Krogan, N. J. *et al.* High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell* **13**, 225–239 (2004).
 19. Guldener, U. *et al.* CYGD: the comprehensive yeast genome database. *Nucleic Acids Res.* **33**, D364–D368 (2005).
 20. Hittinger, C. T., Rokas, A. & Carroll, S. B. Parallel inactivation of multiple *GAL* pathway genes and ecological diversification in yeasts. *Proc. Natl Acad. Sci. USA* **101**, 14144–14149 (2004).
 21. Wolfe, K. H., Morden, C. W. & Palmer, J. D. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl Acad. Sci. USA* **89**, 10648–10652 (1992).
 22. Greig, D., Borts, R. H., Louis, E. J. & Travisano, M. Epistasis and hybrid sterility in *Saccharomyces*. *Proc. R. Soc. Lond. B* **269**, 1167–1171 (2002).
 23. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101**, 9903–9908 (2004).
 24. Postlethwait, J., Amores, A., Cresko, W., Singer, A. & Yan, Y. L. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet.* **20**, 481–490 (2004).
 25. Maere, S. *et al.* Modeling gene and genome duplications in eukaryotes. *Proc. Natl Acad. Sci. USA* **102**, 5454–5459 (2005).
- Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.
- Acknowledgements** We are grateful to D. Bradley, G. Conant, J. Conery, B. Cusack, H. Fraser, N. Khaldi, W.-H. Li, A. Lloyd, A. McLysaght, G. Singer, R. Vilgalys and M. Woolfit for discussion. This study was supported by Science Foundation Ireland.
- Author Contributions** S.W. and K.H.W. did pilot studies. K.P.B. developed the YGOB and the scoring scheme. D.R.S. did all analyses except those of gene function (K.P.B.) and chromosomal rearrangements (J.L.G.). Manual editing of data in YGOB was done equally by all authors. D.R.S. and K.H.W. wrote the manuscript.
- Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to K.H.W. (khwolfe@tcd.ie).

LETTERS

The finished DNA sequence of human chromosome 12

Steven E. Scherer¹, Donna M. Muzny¹, Christian J. Buhay¹, Rui Chen¹, Andrew Cree¹, Yan Ding¹, Shannon Dugan-Rocha¹, Rachel Gill¹, Preethi Gunaratne¹, R. Alan Harris¹, Alicia C. Hawes¹, Judith Hernandez¹, Anne V. Hodgson¹, Jennifer Hume¹, Andrew Jackson¹, Ziad Mohid Khan¹, Christie Kovar-Smith¹, Lora R. Lewis¹, Ryan J. Lozado¹, Michael L. Metzker¹, Aleksandar Milosavljevic¹, George R. Miner¹, Kate T. Montgomery², Margaret B. Morgan¹, Lynne V. Nazareth¹, Graham Scott¹, Erica Sodergren¹, Xing-Zhi Song¹, David Steffen¹, Ruth C. Lovering³, David A. Wheeler¹, Kim C. Worley¹, Yi Yuan¹, Zhengdong Zhang¹, Charles Q. Adams¹, M. Ali Ansari-Lari¹, Mulu Ayele¹, Mary J. Brown¹, Guan Chen¹, Zhijian Chen¹, Kerstin P. Clerc-Blankenburg¹, Clay Davis¹, Oliver Delgado¹, Huyen H. Dinh¹, Heather Draper¹, Manuel L. Gonzalez-Garay¹, Paul Havlak¹, Laronda R. Jackson¹, Leni S. Jacob¹, Susan H. Kelly¹, Li Li², Zhangwan Li¹, Jing Liu¹, Wen Liu¹, Jing Lu¹, Manjula Maheshwari¹, Bao-Viet Nguyen¹, Geoffrey O. Okwuonu¹, Shiran Pasternak¹, Lesette M. Perez¹, Farah J. H. Plopper¹, Jireh Santibanez¹, Hua Shen¹, Paul E. Tabor¹, Daniel Verduzco¹, Lenée Waldron¹, Qiaoyan Wang¹, Gabrielle A. Williams¹, JingKun Zhang¹, Jianling Zhou¹, Baylor College of Medicine Human Genome Sequencing Center Sequence Production Team*, David Nelson¹, Raju Kucherlapati², George Weinstock¹ & Richard A. Gibbs¹

Human chromosome 12 contains more than 1,400 coding genes¹ and 487 loci that have been directly implicated in human disease². The q arm of chromosome 12 contains one of the largest blocks of linkage disequilibrium found in the human genome³. Here we present the finished sequence of human chromosome 12, which has been finished to high quality and spans approximately 132 megabases, representing ~4.5% of the human genome. Alignment of the human chromosome 12 sequence across vertebrates reveals the origin of individual segments in chicken, and a unique history of rearrangement through rodent and primate lineages. The rate of base substitutions in recent evolutionary history shows an overall slowing in hominids compared with primates and rodents.

Among the human chromosome sequencing projects, the chromosome 12 sequencing effort benefited most from an earlier advanced sequence tagged site (STS) physical map, which contained 5,300 large-insert clones and 3,100 markers with an average resolution of 44 kilobases (kb)⁴. After integration with the whole genome fingerprint map⁵, a final tiling path of 1,168 large-insert clones was chosen for sequencing using the clone-by-clone shotgun sequencing strategy. Each clone was finished to community standards (<http://www.genome.gov/10001812>) yielding 130,683,379 base pairs (bp) of nonoverlapping sequence, independently measured as greater than 99.99% accurate⁶. The features and annotations presented here may be viewed as user-specified tracks within the Genboree genome browser (<http://www.genboree.org/Hs.chr12>).

The finished sequence contains just five euchromatic gaps, estimated to total 380 kb by fibre fluorescence *in situ* hybridization (FISH) (C. Wagner-McPherson, personal communication) and by comparison to primate and rodent genome assemblies (Supplementary Table 1). The data extend to 16 and 60 kb from the telomere

terminus repeats at the p and q arms, respectively (H. Riethman, personal communication; <http://www.wistar.upenn.edu/Riethman/>)⁷. The pericentromeric sequences on the p and q arms contain approximately 425 and 600 kb of tandem alpha-satellite repeats, respectively. The alpha satellites do not demonstrate the higher-order structure indicative of the 'core centromere' on either arm, but previously established markers⁸, which are present in BAC clones flanking the centromere, result in a calculated centromere length of 1.395 megabases (Mb) and an overall chromosome length of 132,449,811 bp.

Starting with an automated analysis of human build 33 using the Ensembl pipeline⁹, we manually annotated the chromosome 12 finished sequence using evidence from all publicly available protein and complementary DNA databases, as well as spliced expressed sequence tags (ESTs). We identified a further 282 gene structures beyond the automated output, resulting in a total of 1,435 loci. Using annotation standards developed by the Human Annotation Workshop (<http://www.sanger.ac.uk/HGP/havana/hawk.shtml>), the loci were categorized as 1,294 'known genes', 12 'novel coding sequences (CDS)', 34 'novel transcripts', 2 'putative genes' and 93 'pseudogenes'.

We found 4,427 paralogous pairings to genes on chromosome 12, of which 528 were intrachromosomal. Notably, the density of paralogues correlates well with the density of SINEs (short interspersed elements) and breakpoints observed between the human chromosome and its syntenic regions in avian, rodent and canine genomes (see Fig. 1). Excluding pseudogenes, the average gene density on chromosome 12 is 11.0 genes per Mb, which is relatively typical when compared to the gene densities of other chromosomes. A total of just 11 out of 1,264 RefSeq genes¹ are completely or partially absent from genome build 33, while there are only three partially missing genes (NM_001733, *CIR*; NM_018711, *SVOP*; and NM_020993, *BCL7A*) from build 35.

¹Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. ²Harvard Medical School-Partners Healthcare Center for Genetics and Genomics, Boston, Massachusetts 02115, USA. ³HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK.

*A list of participants and affiliations appears at the end of the paper.

Approximately 58.3% of chromosome 12 genes expressed alternative transcripts, averaging 2.89 transcripts per gene, but ranging as high as 20 (for *UBC*). The majority of these produced altered protein products (2,923 different proteins from among 3,148 alternative transcripts). There were at least 677 partial transcripts, based pri-

marily on spliced EST data, for which we could not identify the complete coding sequence. Therefore, these estimates represent a lower bound on the total alternative splicing activity on the chromosome.

The density of genes across chromosome 12 varies quite widely



Figure 1 | Correlation of syntenic breakpoints with general chromosome landscape features. Tracks are numbered on the left, and syntenic alignments across human chromosome 12 are shown in the top five tracks: 1, human–chicken; 2, human–rat; 3, human–mouse; 4, human–dog; and 5, human–chimpanzee. The inter- and intrachromosomal breakpoints are represented by red and blue gaps, respectively. Aqua gaps indicate regions without sequence alignment, and the centromere is denoted by the grey bar running through all tracks. Purple brackets portray sequence inversions. The density of recent segmental intra- and interchromosomal duplications

from low-copy repeats are shown in tracks 6 and 7, respectively. The incidence of major interspersed (high-copy) repeats is depicted in tracks 8 (LINEs), 9 (LTRs) and 10 (SINEs). The variations in G+C content, and densities of CpG islands, genes and pseudogenes, appear in tracks 11, 12, 13 and 14, respectively, while gene paralogue density, gene density and gene variant density appear in tracks 15, 16 and 17, respectively. Gene density in track 13 is from UCSC 'known genes', and in track 16 is from the nonredundant locus annotations performed in this study. The $y_{\max}$ values in tracks 14–17 reflect the maximum y -axis values obtained for those tracks.

A comparison of physical and genetic maps revealed wide variation in recombination activity, with a slightly higher overall recombination rate in females as expected, and an average of 1.3 centimorgans (cM) per Mb (see Supplementary Fig. 1). There are, however, no extensive recombination ‘deserts’ or ‘jungles’ (with ‘jungles’ defined as having a recombination rate of >3 cM per Mb) as previously described¹⁰.

Alignment to rodent chromosomes (>80 Myr evolutionary separation) shows that approximately 9 major rearrangements have occurred between human chromosome 12 and the hypothetical common ancestor with rat and mouse. The breakpoints identified

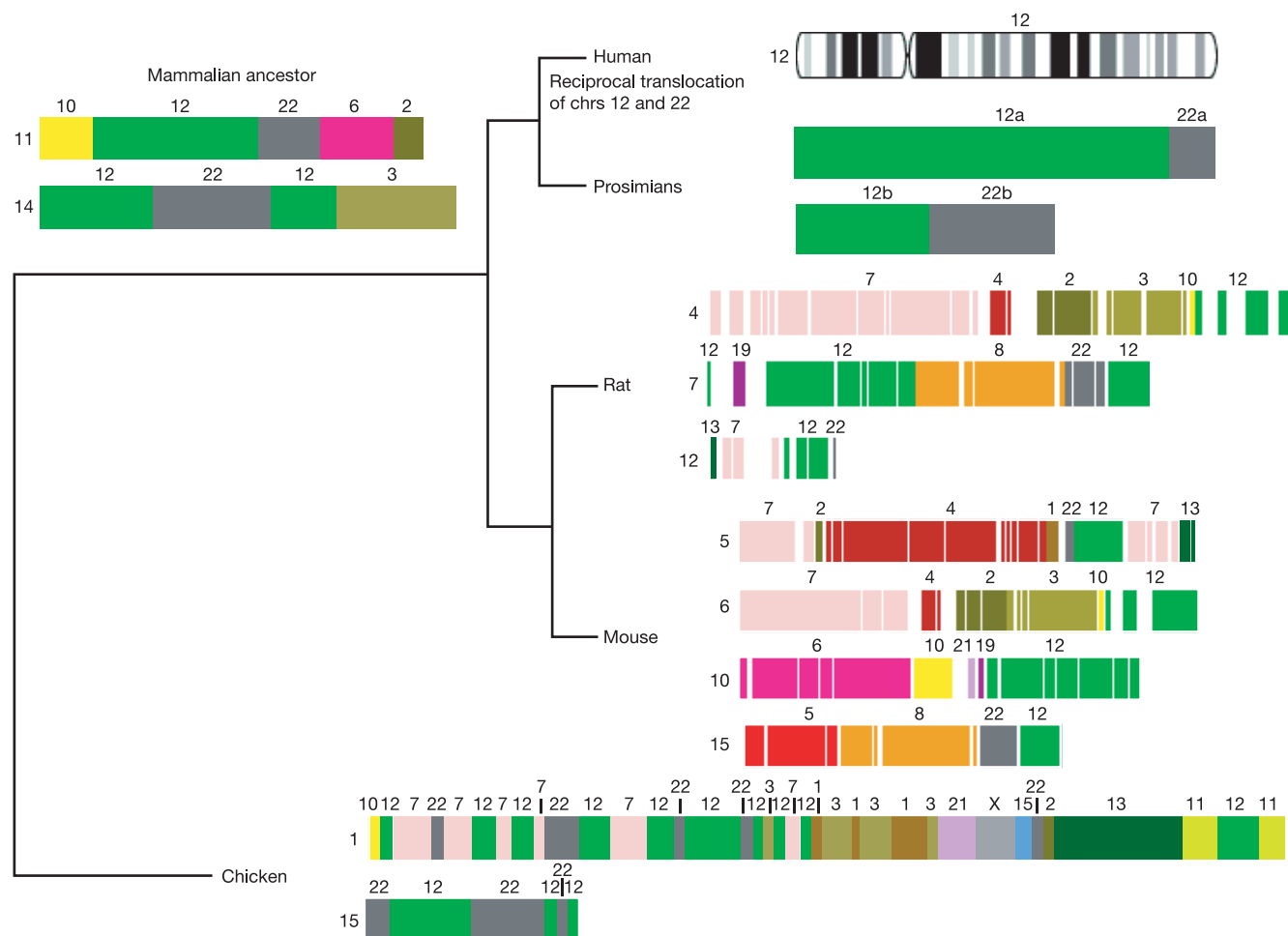


Figure 2 | Human chromosome 12 evolution. Numbers above the chromosome indicate the human orthologous chromosome, and numbers to the left of the chromosome refer to the chromosome of the compared organism. Regions orthologous to human chromosome 12 are shown in green. The rodent comparisons to human were performed using Pash²⁸, and

the ancestral mammal and chicken are adapted from GRIMM (Genome Rearrangements In Man and Mouse)-Synteny computations presented in ref. 30. On the basis of cytogenetic evidence, a double reciprocal translocation of human chromosomes 12 and 22 occurred after the divergence of prosimian primates, and is present in all anthropoid primates.

Table 1 | Human- and chimpanzee-specific base-pair substitution rates

	Intergenic	Intronic	UTR	CDS	Substitutions per year* (non-CDS)
Human	0.0060	0.0056	0.0043	0.0024	9.70×10^{-10}
Chimpanzee	0.0074	0.0069	0.0069	0.0034	1.19×10^{-9}

Base-pair substitution rates were calculated per site within the genomic regions indicated.
 *Substitutions per year was calculated assuming a human–chimpanzee divergence at 6 Myr ago.

between the human and rodent genomes are summarized in Fig. 1. An unexpected feature is the pattern of increased activity at the ends of the chromosome, with a substantial portion of the proximal q arm relatively unaffected by gross rearrangement across mammals. The detailed examination of breakpoints between the available genome sequences does not suggest any obvious sequence features that correlate with the evolutionary change, and the pattern of breaks otherwise generally conforms to that predicted from the known evolutionary relationships. There is an observed increase in the number of intra- and interchromosomal duplications in the regions where breaks occur, but a full analysis awaits a comprehensive description of duplication in the nonhuman species, which are all currently at draft coverage.

There are innumerable ‘small-scale’ differences between regions of the rodent and human chromosomes with conserved synteny. The data presented here confirm earlier studies¹⁴ showing that the rate of change of single base differences and small insertions was greater in the rodent than in hominid lineages, while the deletion rates were slightly lower. Several studies have attempted to correlate the frequency and distribution of these small-scale changes with other genome features. For example, the relative increase in insertions correlates with the active expansion of the overall size of human chromosomes, and a genome-wide burst of SINE insertions (mainly AluJ and AluS repeats) in the human lineage. Furthermore, the density of SINEs shows a modest positive correlation with the microinsertion rate ($r = 0.66$), a strong negative correlation with the substitution rate ($r = -0.73$), but no correlation with micro-deletions. The SINE distribution correlations may be due to local G+C content, but the relationships are complicated. For example, G+C density shows modest-to-strong positive correlation with both SINEs and insertions. However, SINEs are known to prefer to insert in locally G+C-rich DNA, and this alone may govern the connection between G+C content and the other events. The human chromosomes have a lower G+C content than in rodents, leading to a generally lower rate of G or C gain substitutions (compared with A/T gains) (Supplementary Table 3) exacerbating the complexity of the effect of G+C SINE insertions^{14,15}.

To better understand more recent small-scale changes leading to the current human sequence, we compared chromosome 12 to the

available orthologous regions of the chimpanzee and rhesus macaque genomes. Although neither nonhuman assembly represents a ‘deep draft’, high sequence identity enabled alignment of about 87% of the human chromosome to both nonhuman primates. The availability of two additional primate genomes allowed reconstruction of the ancestral hominoid genome from the three-way alignment. Human- and chimpanzee-specific evolutionary changes, listed in Table 1, were defined by differences in each species compared with the ancestral hominoid.

Previous comparisons of substitution rates per year in neutral sites between human, rat and mouse¹⁴ revealed a lower rate in the lineage from the human–rodent ancestor to human (1.73×10^{-9}) compared with the rate from the human–rodent ancestor to rat (4.95×10^{-9}) and mouse (4.84×10^{-9}). Other findings indicate hominoid substitution rates are lower than the rates in Old World monkeys and in other eutherian mammals^{16–18}. Our analysis reveals further slow down in the branch from the hominoid ancestor to human (9.70×10^{-10}) compared with chimpanzee (1.19×10^{-9}). (See note added in proof.)

The pattern of insertion–deletions (indels) in human chromosome 12 relative to the hominoid ancestor is summarized in Table 2. Microsatellite expansion events, predominantly A/T mononucleotide and CA dinucleotide expansions, comprise 19% of insertion events in the 1–100-bp range. Trinucleotide expansions are more common in coding (1.62 per million base pairs (Mbp)) than noncoding regions (0.727 per Mbp). Comparisons of insertions in the 101–8,000-bp range to the “Retroposed Genes” track of the University of California at Santa Cruz (UCSC) Genome Browser (<http://genome.ucsc.edu/>; ref. 19) revealed 12 human-specific pseudogenes. Comparisons to known segmental duplications²⁰ revealed nine human-specific duplications. Human-specific retrotransposon insertions in the 101–8,000-bp size range are classified by type in Table 3. Consistent with previous results²¹, SINE insertion rates in the human lineage are twofold higher than in the chimpanzee lineage. All retrotransposon insertions were intergenic or intronic, except for an AluY insertion in the 3′ untranslated region (UTR) of *RAB21*, a member of the *RAS* oncogene family. The overall ratio of inserted to deleted base pairs was 1.67, lengthening the human chromosome by 2.69% relative to the ancestral hominoid and consistent with previously detected evolutionary increases in the size of the human genome²².

Chromosome 12 is rich in disease-associated loci, and a total of 487 disease genes have been assigned to this chromosome², accounting for 5.2% of all disease genes. In Supplementary Tables 4 and 5, we present the most cited genes on chromosome 12 together with the medically relevant genes sorted by function, respectively.

With regard to disease, chromosome 12 is best known for its links with cancer. Three genes mapping to the chromosome have been associated with cancer-related chromosome translocations. The most prominent of these is the *ETV6* (*TEL1*) gene at 12p13,

Table 2 | Human-specific insertion and deletion rates

	Intergenic	Intronic	UTR	CDS
1–100-bp indels				
Insertion events	0.3636	0.3938	0.2748	0.0316
Inserted bases	1.9768	2.0958	1.2722	0.2662
Microsatellite expansion events	0.0669	0.0779	0.0344	0.0024
Microsatellite expansion bases	0.4483	0.4512	0.0022	0.0002
Deletion events	0.3659	0.3672	0.3092	0.0267
Deleted bases	2.5017	2.4870	1.5683	0.2007
101–8,000-bp indels				
Insertion events	0.0131	0.0173	0.0059	0.0049
Inserted bases	9.0789	12.9904	9.3402	2.1110
Retrotransposon insertion events	0.0061	0.0075	0.0012	0.0000
Retrotransposon insertion bases	4.8921	5.4302	0.3897	0.0000
Deletion events	0.0235	0.0206	0.0154	0.0000
Deleted bases	5.6285	4.7809	5.1019	0.0000

Indels were calculated per kilobase of sequence within the genomic regions indicated.

Table 3 | Human-specific insertions of retrotransposons

	Insertion events		Inserted bases per Mb
	Total count	Per Mb	
SINEs	419	3.72	1083.07
AluY	338	3.00	358.90
AluYa5	116	1.03	316.44
AluYb8	79	0.70	220.21
LINEs	228	2.02	3010.20
L1	225	2.00	3002.83
L1PA2	52	0.46	1494.21
L1HS	13	0.12	312.74
LTRs	51	0.45	328.00
SVAs*	50	0.44	628.13

*SVAs are hominoid-specific composite retrotransposons comprised of SINE-R, VNTR (variable number of tandem repeats) and Alu sequences.

encoding an ETS-like transcription factor, which has a central role in haematopoietic malignancies including acute lymphoblastic leukaemia (ALL), acute myeloblastic leukaemia (AML) and myelodysplastic syndromes (MDS). Furthermore, 5% of children with ALL have 12p13–p12 deletions²³. The other two genes on this chromosome involved in oncogenic gene fusions are *DDIT3* (or *CHOP*; 12q13.1), often rearranged in myxoid liposarcoma, and *HMG2* (12q15), which has been found fused to various genes in lipoma, salivary adenoma, uterine leiomyoma and multiple lipomatosis. In addition, several genes mapped to this chromosome have been directly or indirectly associated with cancer, including *BCL7A* (12q24.13) in B-cell non-Hodgkin lymphoma, *P2RX7* (12q24) in susceptibility to chronic lymphatic leukaemia (CLL), *YEATS4* (12q13–q15) in glioma, *CDK4* (12q14) in melanoma, *ACVR1B* (12q13) in pancreatic cancer, and *KRAS* (12p12.1) in colorectal adenoma.

There are 16 genes associated with movement disorders mapped to chromosome 12, as well as two diabetes-related genes, genes associated with four separate mood disorders, and four genes involved in heart disease. Of the 28 members of the keratin gene family mapped to this chromosome, 12 have been linked to skin and hair disorders. Chromosome 12 also harbours the human *CD4* locus, which encodes the main receptor for the human immunodeficiency virus, at 12p3.1. This locus is within one of the most dense gene clusters in the genome, and has been the focus of many functional and population genetic studies²⁴.

A primary goal in generating a high quality, finished reference sequence for chromosome 12 is to provide the research community with the resources to accelerate the search for additional disease-causing genes. As an example, Li and co-workers²⁵ were recently able to demonstrate an association between sequence variants of the *GAPDH* gene on chromosome 12 and late-onset Alzheimer's disease (LOAD).

The finished sequence of chromosome 12 reported here marks the beginning of more extensive studies aimed at understanding our evolutionary history, together with the variation in genome structure and sequence that defines us as individuals.

Note added in proof: Recent data³¹ showed that of the nine human chromosomes assayed, chromosome 12 demonstrated the greatest slow down in single nucleotide substitution rate.

METHODS

Mapping and sequencing. Gap closure, sequencing and finishing strategies, and clone integrity assays, are described in Supplementary Methods. BAC clone overlaps were verified by BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>) in-house using a locally installed copy of the software, and polymorphic regions within overlaps were confirmed by polymerase chain reaction (PCR) using a bi-gender, multi-ethnic pool of genomic DNA isolated from eight individuals (J. Belmont, personal communication). Coverage, integrity and clone order were analysed using available genetic and radiation hybrid map markers (see Supplementary Methods and Supplementary Fig. 2). Additionally, we aligned the unique paired-end fosmid reads to the finished chromosome sequence, and concluded that the limited set of clustered size discrepancies reflected probable polymorphisms between clone libraries. The unique paired fosmid end-sequence (Broad Institute) analysis was performed using sequences downloaded from the UCSC Genome Browser that were checked for both paired end-sequence orientation and resulting insert size.

Annotation. Manual curation identified each known gene, novel gene and novel transcript locus, defined as a set of one or more transcripts that share at least one exon of coding sequence (in frame) and supported by full-length and partial human or vertebrate cDNA sequences having a best-in-genome BLAST or blat (<http://genome.ucsc.edu/cgi-bin/hgBlat>) hit ($\geq 98\%$ identity). All analyses were performed in-house using locally installed copies of the software. The cDNA/RefSeq sequences were compared to the genomic sequence to place exons. All intron/exon splice sites for all predicted exons were examined for canonical splice motifs. Coding regions were examined for a best-fit open reading frame (ORF). The 5' and 3' UTRs were annotated and extended using available EST and cDNA evidence, and poly(A) sites and poly(A) signals were annotated on each gene where identified. Alternative splice variants were identified from cDNA, EST and protein evidence. The translation product for each coding sequence was verified using Swissprot. Pseudogenes were defined as sequences with no direct evidence for expression but that match with a

high score to a spliced messenger RNA or spliced EST from elsewhere in the genome. This is a more stringent definition than has been applied by others in broad genomic screens of pseudogenes, and results in a fivefold-lower count across chromosome 12 than previously reported²⁶. For paralogue analysis, protein sequences corresponding to the 'known genes' track of the UCSC Genome Browser²⁷ were compared in an all-against-all BLAST search. Two loci were defined as paralogues if there was a match of any of their transcript variants with the following criteria: expect value cutoff of 10^{-10} or less, the lengths of the matching transcripts were within 20% of each other, and the match length extended over 70% of the average length of the two sequences. The complete set of annotations has been submitted to the Vega database (http://vega.sanger.ac.uk/Homo_Sapiens/).

Landscape features. For CpG analysis, a CpG island has been defined as an expanse of >200 nucleotides in which the G+C content is greater than 50% and the ratio of observed CG dinucleotides to expected in the segment is >0.6 . We developed databases of known microRNAs, small nuclear (sn) RNAs and small nucleolar (sno) RNAs, and used tRNAscan-SE v.1.23 (<http://www.genetics.wustl.edu/eddy/tRNAscan-SE>; installed locally) and BLAST to search for non-coding RNA sequences. We identified recent intra- and interchromosomal segmental duplications by BLAST searching the repeat-masked chromosome sequence against itself and the rest of the human genome. The duplication densities were calculated by averaging the duplications of each base over nonoverlapping 100-kb windows after filtering low identity matches ($<90\%$). The densities of SINEs, LINEs and long terminal repeats (LTRs) were calculated from repeat-masked data using 100-kb windows. The G+C density was calculated by counting the G+C content over nonoverlapping 100-kb windows. The densities of CpG islands (UCSC), genes (Baylor College of Medicine Human Genome Sequencing Center annotations), and pseudogenes (as defined above) were counted and displayed using 1-Mb windows. Markers were placed on the genomic sequence using a combination of locally installed e-PCR (<http://www.ncbi.nlm.nih.gov/sutils/e-pcr/>) and BLAST software packages.

Comparative analysis. The multiple alignments of human, chimp (*panTro1*), dog (*canFam1*), mouse (*mm5*), rat (*rn3*), chicken (*galGal2*), zebrafish (*danRer1*) and fugu (*fr1*) were downloaded from the UCSC Genome Browser (<http://hgdownload.cse.ucsc.edu/>; May 2004/hg17-Build35). The pairwise synteny blocks between human and other species were parsed out with Synteny-Parser (X. Song & G. Weinstock, unpublished perl script), which was tuned to include all visible chromosome rearrangements in the dot plot. Rhesus scaffolds from the Mmul_0.1 preliminary assembly were mapped to human using Pash²⁸. Rhesus scaffolds that mapped to human chromosome 12 by both Pash and the human–rhesus Alignment Net (UCSC) were aligned with orthologous human regions and chimpanzee regions from the Human–Chimpanzee Reciprocal-Best Chain alignments (UCSC) using MLAGAN²⁹.

Received 17 December; accepted 31 December 2005.

- Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
- Online Mendelian Inheritance in Man (OMIM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, Maryland) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, Maryland) (<http://www.ncbi.nlm.nih.gov/omim/>) (2000).
- Yu, F. et al. Positive selection of a pre-expansion CAG repeat of the human *SCA2* gene. *PLoS Genet.* **1**, e41 (2005).
- Montgomery, K. T. et al. A high-resolution map of human chromosome 12. *Nature* **409**, 945–946 (2001).
- McPherson, J. D. et al. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
- Schmutz, J. et al. Quality assessment of the human genome sequence. *Nature* **429**, 365–368 (2004).
- Knight, S. J. et al. An optimized set of human telomere clones for studying telomere integrity and architecture. *Am. J. Hum. Genet.* **67**, 320–332 (2000).
- Vermeech, J. R. et al. A physical map of the chromosome 12 centromere. *Cytogenet. Genome Res.* **103**, 63–73 (2003).
- Curwen, V. et al. The Ensembl automatic gene annotation system. *Genome Res.* **14**, 942–950 (2004).
- Yu, A. et al. Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
- Wienberg, J. Fluorescence *in situ* hybridization to chromosomes as a tool to understand human and primate genome evolution. *Cytogenet. Genome Res.* **108**, 139–160 (2005).
- Nickerson, E. & Nelson, D. L. Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics* **50**, 368–372 (1998).
- Kehrer-Sawatzki, H., Sandig, C. A., Goidts, V. & Hameister, H. Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and

- the homologous chromosome 12 in humans. *Cytogenet. Genome Res.* **108**, 91–97 (2005).
14. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
 15. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
 16. Li, W. H. & Tanimura, M. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326**, 93–96 (1987).
 17. Steiper, M. E., Young, N. M. & Sukarna, T. Y. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid–cercopithecoid divergence. *Proc. Natl Acad. Sci. USA* **101**, 17021–17026 (2004).
 18. Yi, S., Ellsworth, D. L. & Li, W. H. Slow molecular clocks in Old World monkeys, apes, and humans. *Mol. Biol. Evol.* **19**, 2191–2198 (2002).
 19. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100**, 11484–11489 (2003).
 20. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11**, 1005–1017 (2001).
 21. Hedges, D. J. *et al.* Differential *Alu* mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* **14**, 1068–1075 (2004).
 22. Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13**, 358–368 (2003).
 23. Stegmaier, K. *et al.* Frequent loss of heterozygosity at the *TEL* gene locus in acute lymphoblastic leukemia of childhood. *Blood* **86**, 38–44 (1995).
 24. Ansari-Lari, M. A. *et al.* A gene-rich cluster between the CD4 and triosephosphate isomerase genes at human chromosome 12p13. *Genome Res.* **6**, 314–326 (1996).
 25. Li, Y. *et al.* Association of late-onset Alzheimer's disease with genetic variation in multiple members of the *GAPD* gene family. *Proc. Natl Acad. Sci. USA* **101**, 15688–15693 (2004).
 26. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13**, 2541–2558 (2003).
 27. Hsu, F. *et al.* The UCSC Proteome Browser. *Nucleic Acids Res.* **33** (suppl. 1), D454–D458 (2005).
 28. Kalafus, K. J., Jackson, A. R. & Milosavljevic, A. Pash: Efficient genome-scale sequence anchoring by Positional Hashing. *Genome Res.* **14**, 672–678 (2004).
 29. Brudno, M. *et al.* LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731 (2003).
 30. Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A. & Tesler, G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.* **15**, 98–110 (2005).
 31. Elango, N., Thomas, J. W., NISC Comparative Sequencing Program & Soojin, V. Y. Variable molecular clocks in hominoids. *Proc. Natl Acad. Sci. USA* **103**, 1370–1375 (2006).
- Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.
- Acknowledgements** This work was supported by NIH grants to R.K. and to R.A.G. We acknowledge and thank the genome sequencing community for generating the data sets used in our comparative analysis. We also acknowledge the following members of the HUGO Gene Nomenclature Committee: S. Povey (chair), E. A. Bruford, V. K. Khodiyar, M. J. Lush, K. M. B. Sneddon, T. P. Sneddon, C. C. Talbot Jr and M. W. Wright.
- Author Information** The chromosome 12 sequence has been deposited into GenBank under the accession number NC_000012. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.S. (sscherer@bcm.tmc.edu).

Baylor College of Medicine Human Genome Sequencing Center Sequence Production Team Carlana C. Allen¹, Anita G. Amin¹, Vivian Anyalebechi¹, Michael Bailey¹, Joseph A. Barbaria¹, Kesha E. Bimage¹, Nathaniel P. Bryant¹, Paula E. Burch¹, Carrie E. Burkett¹, Kevin L. Burrell¹, Eliana Calderon¹, Veronica Cardenas¹, Kelvin Carter¹, Kristal Casias¹, Iracema Cavazos¹, Sandra R. Cavazos¹, Heather Ceasar¹, Joseph Chacko¹, Sheryl N. Chan¹, Dean Chavez¹, Constantine Christopoulos¹, Joseph Chu¹, Raynard Cockrell¹, Caroline D. Cox¹, Michelle Dang¹, Stephanie R. Dathorne¹, Robert David¹, Candi Mon'Et Davis¹, Latarsha Davy-Carroll¹, Denise R. Deshazo¹, Jeremy E. Donlin¹, Lisa D'Souza¹, Kristy A. Eaves¹, Amy Egan¹, Alexandra J. Emery-Cohen¹, Michael Escotto¹, Nicole Flagg¹, Lisa D. Forbes¹, Abdul M. Gabisi¹, Melissa Garza¹, Cerissa Hamilton¹, Nicholas Henderson¹, Omar Hernandez¹, Sandra Hines¹, Marilyn E. Hogues¹, Mei Huang¹, DeVincent G. Idlebird¹, Rudy Johnson¹, Angela Jolivet¹, Sally Jones¹, Ryan Kagan¹, Laquisha M. King¹, Belita Leal¹, Heather Lebow¹, Sandra Lee¹, Jaclyn M. LeVan¹, Lakeshia C. Lewis¹, Pamela London¹, Lorna M. Lorensuhewa¹, Hermela Loulseged¹, Demetria A. Lovett¹, Alice Lucier¹, Raymond L. Lucier¹, Jie Ma¹, Renita C. Madu¹, Patricia Mapua¹, Ashley D. Martindale¹, Evangelina Martinez¹, Elizabeth Massey¹, Samantha Mawhiney¹, Michael G. Meador¹, Sylvia Mendez¹, Christian Mercado¹, Iracema C. Mercado¹, Christina E. Merritt¹, Zachary L. Miner¹, Emmanuel Minja¹, Teresa Mitchell¹, Farida Mohabbat¹, Khatera Mohabbat¹, Baize Montgomery¹, Niki Moore¹, Sidney Morris¹, Mala Munidas¹, Robin N. Ngo¹, Ngoc B. Nguyen¹, Elizabeth Nickerson¹, Ogechi O. Nwaokeme¹, Stanley Nwokenkwo¹, Melissa Obregon¹, Maryann Oguh¹, Njideka Oragunye¹, Rodolfo J. Oviedo¹, Bridgette J. Parish¹, David N. Parker¹, Julia Parrish¹, Kenya L. Parks¹, Heidie A. Paul¹, Brett A. Payton¹, Agapito Perez¹, William Perrin¹, Adam Pickens¹, Eltrick L. Primus¹, Ling-Ling Pu¹, Maria Puazo¹, Miyo M. Quiles¹, Juana B. Quiroz¹, Dina Rabata¹, Kacy Reeves¹, San Juana Ruiz¹, Hongmei Shao¹, Ida Sisson¹, Titilola Sonaike¹, Richard P. Sorelle¹, Angelica E. Sutton¹, Amanda F. Svatek¹, Leah Anne Svetz¹, Kavitha S. Tamerisa¹, Tineace R. Taylor¹, Brian Teague¹, Nicole Thomas¹, Rachel D. Thorn¹, Zulma Y. Trejos¹, Brenda K. Trevino¹, Ogechi N. Ukegbu¹, Jeremy B. Urban¹, Lydia I. Vasquez¹, Virginia A. Vera¹, Donna M. Villasana¹, Ling Wang¹, Stephanie Ward-Moore¹, James T. Warren¹, Xuehong Wei¹, Flower White¹, Angela L. Williamson¹, Regina Wleczyk¹, Hailey S. Wooden¹, Steven H. Wooden¹, Jennifer Yen¹, Lillienne Yoon¹, Vivienne Yoon¹ & Sara E. Zorrilla¹

Affiliation for participants: ¹Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.

LETTERS

A specific amyloid- β protein assembly in the brain impairs memory

Sylvain Lesné¹, Ming Teng Koh⁴, Linda Kotilinek¹, Rakez Kayed⁶, Charles G. Glabe⁶, Austin Yang⁷, Michela Gallagher⁴ & Karen H. Ashe^{1,2,3,5}

Memory function often declines with age¹, and is believed to deteriorate initially because of changes in synaptic function rather than loss of neurons². Some individuals then go on to develop Alzheimer's disease with neurodegeneration. Here we use *Tg2576* mice, which express a human amyloid- β precursor protein (APP) variant linked to Alzheimer's disease, to investigate the cause of memory decline in the absence of neurodegeneration or amyloid- β protein amyloidosis. Young *Tg2576* mice (<6 months old) have normal memory and lack neuropathology, middle-aged mice (6–14 months old) develop memory deficits without neuronal loss, and old mice (>14 months old) form abundant neuritic plaques containing amyloid- β (refs 3–6). We found that memory deficits in middle-aged *Tg2576* mice are caused by the extracellular accumulation of a 56-kDa soluble amyloid- β assembly, which we term A β *56 (A β star 56). A β *56 purified from the brains of impaired *Tg2576* mice disrupts memory when administered to young rats. We propose that A β *56 impairs memory independently of plaques or neuronal loss, and may contribute to cognitive deficits associated with Alzheimer's disease.

Poor memory function can predict Alzheimer's disease up to 15 years before diagnosis⁷, and non-demented individuals at risk genetically for Alzheimer's disease show abnormalities in functional brain imaging tests^{8,9}. These and other studies imply that Alzheimer's disease has an insidious onset, which blurs the boundary between age-associated memory impairment and Alzheimer's disease¹⁰. *Tg(APP^{SWE})2576Kahs* mice (hereafter *Tg2576* mice) express a human amyloid- β precursor protein (APP) variant linked to Alzheimer's disease, and develop many neuropathological features of Alzheimer's, including amyloid plaques, dystrophic neurites and inflammatory changes^{3,4,11}. However, *Tg2576* mice lack neurofibrillary tangles, significant neuronal loss and gross atrophy⁴. They may therefore be a good model to study pre-clinical stages of Alzheimer's disease, before the diagnosis of dementia or the onset of neuronal loss.

In *Tg2576* mice, as in other APP transgenic mice, there is strong evidence that amyloid- β (A β) is responsible for age-related memory decline^{6,12,13}. However, there are several paradoxical findings about the relationship between A β and cognitive decline that suggest a complex role for A β in cognitive impairment. For example, spatial reference memory in *Tg2576* mice declines modestly but significantly at 6 months of age and then remains stable for 7 to 8 months (Fig. 1a, b). However, no candidate A β species measured to date corresponds with the decline in memory observed at 6 months and the cognitive stability observed thereafter (see Supplementary Table 1). Hence, we are faced with the paradox that a rapidly increasing amount of A β , the molecule believed to be responsible

for memory loss, is associated with no change in memory function. One solution to this conundrum is to posit the existence of soluble A β assemblies that disrupt memory^{6,14–16}, which we designated A β * (A β star) and sought to identify in *Tg2576* mice.

A challenge in analysing A β in the brain lies in reliably separating the specific cellular pools of A β (for example, extracellular, intracellular, membrane-associated and insoluble). We overcame this obstacle by developing a high-fidelity extraction procedure that separates proteins in known cellular compartments (Supplementary Fig. 1). Our new extraction method allowed us to quantify and compare four independent pools of transgene-derived A β species.

To resolve the problem of a mismatch between A β levels and memory deficits, we used our extraction procedure to search for A β * in *Tg2576* mice between 4 and 25 months of age. We required candidate A β * molecules to satisfy two criteria. First, their appearance should coincide with memory loss at 6 months. Second, their levels should remain stable in middle-aged mice (6–14 months old). By immunoblotting immunoglobulin-depleted forebrain extracts, we found a set of apparent assemblies of A β in the soluble, extracellular-enriched fraction from 6-month-old mice (Fig. 1c). In addition to a faint 4-kDa band corresponding to A β monomers, 6E10- and 4G8-immunoreactive proteins (see Methods) were detected at molecular masses theoretically corresponding to trimeric (14 kDa), hexameric (27 kDa), nonameric (40 kDa) and dodecameric (56 kDa) A β _{1–42} assemblies. These species represent multiples of trimeric A β oligomers, with high-molecular-mass assemblies (>20 kDa) appearing in mice older than 6 months. The detection of similar bands using 6E10 and 4G8 antibodies excludes the possibility that they represent degradation products of soluble APP, which lacks the mid-domain A β epitope (A β _{17–24}) recognized by 4G8 (Supplementary Fig. 2a). The bands were not recognized by 22C11 or APPC17-Cter antibodies, indicating they were neither APP nor APP cleavage-end products (data not shown).

Although this result suggests that ageing induces A β trimers to associate and form high-molecular-mass assemblies, we also considered the possibility that they might represent A β oligomers complexed to binding proteins. However, this is unlikely on the basis of their biochemical properties and immunospecificity. First, we examined their properties in urea, a common denaturant of globular proteins¹⁷. To our surprise, the A β oligomers were unaltered in SDS–polyacrylamide gel electrophoresis (SDS–PAGE) containing 8 M urea (Fig. 2a). However, when exposed to $\geq 10\%$ hexafluoroisopropanol (HFIP), a solvent with strong hydrogen-bonding properties, the theoretical hexamers, nonamers and dodecamers depolymerized, with a parallel increase in levels of tetramers, trimers and, to a lesser extent, monomers (Fig. 2b). In >20% HFIP, only the

¹Departments of Neurology and ²Neuroscience, and ³Graduate Program in Neuroscience, University of Minnesota Medical School, Minneapolis, Minnesota 55455, USA.

⁴Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, Maryland 21218, USA. ⁵Geriatric Research Education Clinical Center, Minneapolis VA Medical Center, Minneapolis, Minnesota 55417, USA. ⁶Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92697, USA. ⁷Department of Pharmaceutical Sciences, University of Southern California, Los Angeles, California 90089, USA.

putative trimers remained; these dissociated into A β monomers in >55% HFIP (not shown). These data suggest that the high-molecular-mass A β complexes are held together by urea-resistant hydrogen bonds, not by covalent bonds. In addition, the greater resistance of trimers to denaturation supports our conjecture that trimers are the fundamental A β assembly unit *in vivo*. Second, A11 antiserum, which specifically detects soluble amyloid assemblies distinct from fibrillar A β , detected only 27–56 kDa complexes in extracellular-enriched extracts, confirming previous observations that A11 reacts with A β oligomers larger than tetramers¹⁸ (Supplementary Fig. 2b). The immunoreactivity of the high-molecular-mass complexes with 6E10, 4G8 and A11 antibodies indicates that they are soluble A β oligomers. Finally, we examined the native size of soluble A β assemblies under non-denaturing conditions in order to demonstrate that they were not artificially generated during SDS-PAGE. Extracellular-enriched extracts were fractionated by non-denaturing size-exclusion chromatography (SEC) and subjected to SDS-PAGE (Fig. 2c). High- and low-molecular-mass A β assemblies (at expected intervals) were collected in different fractions, arguing against the possibility that SDS or self-oligomerization triggered the formation of the A β oligomers. Collectively, our results suggest that endogenous A β forms a ladder of stable, soluble, physiological assemblies comprised of trimers and multiples of trimers in *Tg2576* mice older than 6 months.

Trimers and hexamers were excluded as components of A β^* ,

because they were present before memory impairment (<6 months). However, a 56-kDa band appeared in extracts at 6 months of age, along with lesser quantities of a 40-kDa species; these correspond to theoretical A β_{1-42} dodecamers and nonamers, respectively (Fig. 1c). The mean levels of the soluble A β assemblies remained stable after 6 months of age, although there was considerable variability between animals of the same age (Supplementary Fig. 2c–e). As the 56-kDa and 40-kDa species appeared at 6 months of age and remained stable between the ages of 6 and 13 months, they fulfilled the criteria for being designated as A β^* and thus represented viable candidates for A β assemblies that cause memory deficits. We found no further increase in A β assemblies in old mice to correspond to the second drop in memory function at 15 months (Fig. 1a, b); by this age it is possible that the abundant plaques with prominent dystrophic neurites disrupt synaptic function sufficiently to cause further memory impairment¹⁹.

The variability in levels of A β assemblies between animals of the same age provided an opportunity to examine correlations between the different A β oligomers and memory impairment. To do this, we compared spatial memory and levels of soluble A β species in two groups of 5- and 6-month-old *Tg2576* mice (Fig. 3). Monomers, trimers and hexamers of A β did not correlate significantly with performance in 5- or 6-month-old mice (Fig. 3a–f). However, we observed significant inverse relationships between the levels of 56-kDa and 40-kDa assemblies and memory performance (Fig. 3g, h), with

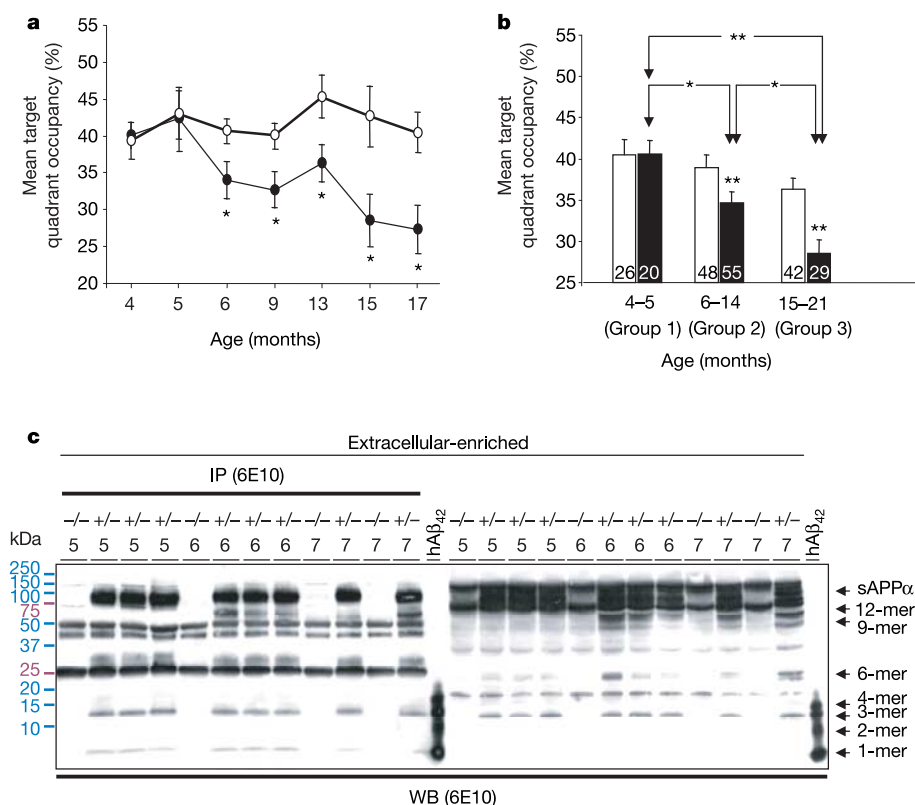


Figure 1 | Temporal patterns of soluble A β oligomers and memory decline in *Tg2576* mice. **a**, Memory retention in 4–17-month-old mice shows a progressive but irregular decline with periods of stability. Filled circles, *Tg2576*^{+/+} mice; open circles, *Tg2576*^{-/-} mice. *Tg2576*^{-/-} and *Tg2576*^{+/+} denote mice harboring no (non-transgenic) or one transgene array, respectively. Asterisk, ANOVA ($P < 0.01$) followed by *t*-test ($P < 0.01$). **b**, Temporal analysis of spatial memory shows three stages of performance. Filled bars, *Tg2576*^{+/+} mice; open bars *Tg2576*^{-/-}. Numbers inside bars denote numbers of mice. Asterisk, $P < 0.01$; two asterisks, $P < 0.001$ (ANOVA). Target quadrant occupancy scores ($\pm$ s.e.m.) during probe trials are a measure of spatial reference memory retention. **c**, Identification of A β

oligomers in soluble, extracellular-enriched extracts of proteins from brains of 5-, 6-, and 7-month-old mice (age indicated above lanes), assessed by western blot (WB) with or without immunoprecipitation (IP). The intensity of the 40-kDa band was $33.92 \pm 12.5\%$ (mean $\pm$ s.e.m.) ($n = 4$) of the intensity of the 56-kDa band. Synthetic human A β_{1-42} peptide (hA β_{42}) was used as a size marker and positive control (right lane). Arrows indicate respective migration positions of monomers (1-mer), dimers (2-mer), trimers (3-mer), tetramers (4-mer), hexamers (6-mer), nonamers (9-mer), dodecamers (12-mer) and sAPP α (secreted form of APP that has been cleaved by α -secretase).

the 56-kDa assembly correlating substantially better than the 40-kDa assembly ($r^2 = 0.66$ and 0.45 , respectively). There were no correlations between the levels of any A β oligomers and performance in the cued phase of a water maze test (see Methods), arguing against their effects on non-cognitive aspects of behaviour (Supplementary Fig. 3).

Because intracellular A β has been proposed to disrupt memory in $3 \times$ Tg-AD transgenic mice (which express APP, presenilin-1 and tau variants)²⁰, we performed additional analyses to evaluate the potential accumulation of A β species within brain cells. Only trimeric and monomeric A β species were detected in the soluble, intracellular-enriched fraction, with no modulation between 5 and 6 months of age (Supplementary Fig. 4a). We also examined full-length

APP and C-terminal fragments (CTFs) in membrane-associated fractions, and found no change in the levels of APP, CTF- β s or CTF- α (Supplementary Fig. 4b, c). In cytosolic extracts of primary Tg2576 cortical neurons cultured from 14–15-day-old embryos, monomers and trimers were the only A β species detected (Supplementary Fig. 5), supporting our *in vivo* findings in young Tg2576 mice. Thus, we found no intracellular or membrane-associated A β species or CTFs correlating with the onset of memory deficits in 6-month-old Tg2576 mice.

The observations that the 56-kDa A β assembly appears at 6 months of age (when memory declines), that its levels are stable in ageing mice, that it is more abundant than the 40-kDa complex (Fig. 1c), and that it shows the strongest correlation with memory

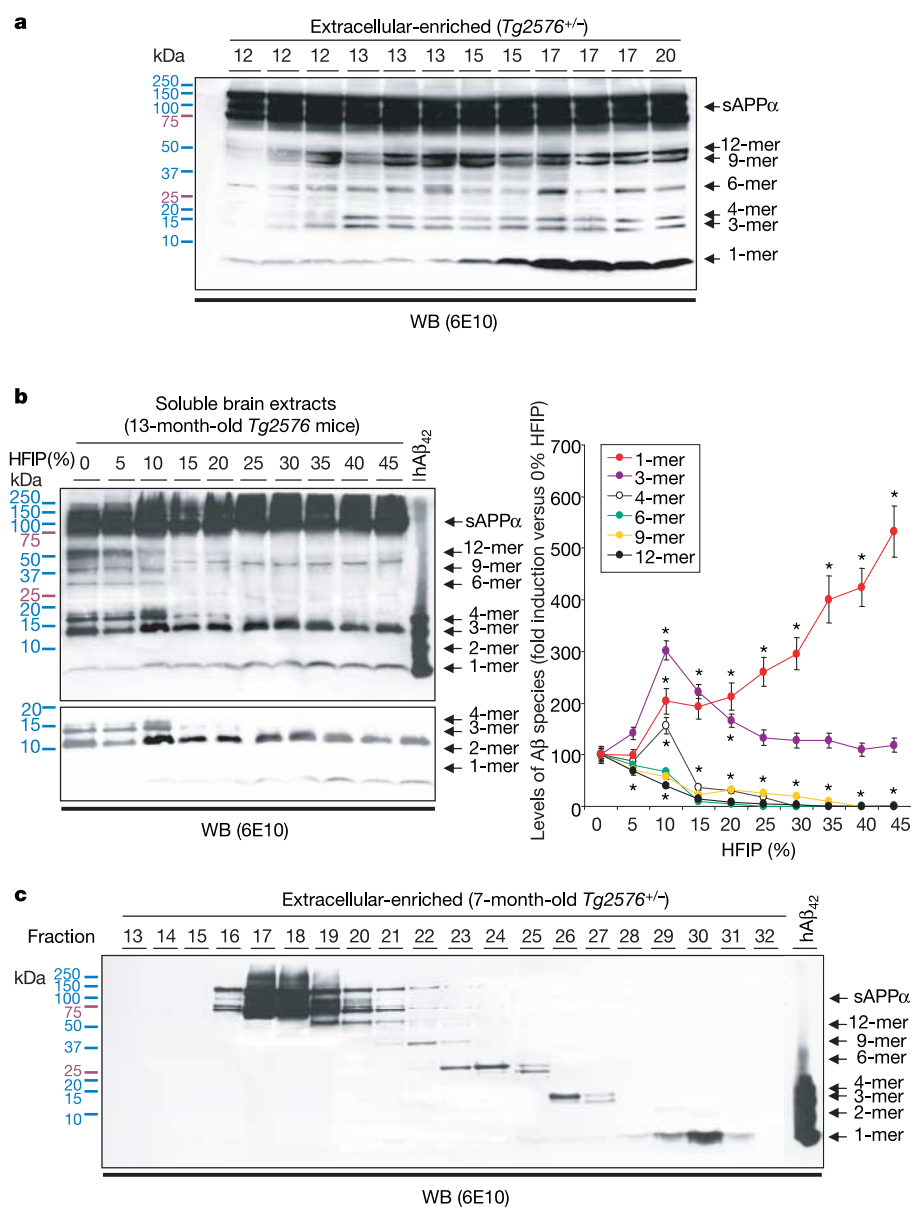


Figure 2 | Biochemical properties of A β assemblies in Tg2576 mice. **a**, The presence of 8 M urea did not alter the electrophoretic pattern of A β oligomers in extracellular-enriched extracts from 12- to 20-month-old brains of Tg2576^{+/-} mice that were probed with 6E10 antibodies. **b**, Soluble high-molecular-mass A β oligomers depolymerize in $\geq 15\%$ HFIP, with concomitant enrichment of both monomeric and trimeric A β (lower exposure provided for enhanced contrast). The data for fold-change in A β species (right panel) is the mean $\pm$ s.e.m. of 3 experiments. **c**, SDS-PAGE analysis of brain extracts from 7-month-old Tg2576^{+/-} mice fractionated by

SDS-free size-exclusion chromatography (SEC) shows that A β oligomers migrate at expected molecular masses. The data confirm that high-molecular-mass A β oligomers are not artificially generated from monomeric or trimeric A β species during electrophoresis, and that trimers are not degradation products of high-molecular-mass oligomers. Bands revealed at 75 and 150 kDa are non-specific bands caused by the modified avidin used for immunoblotting (also present in blots of extracts from non-transgenic mice).

impairment suggest that it is the most likely candidate for A β^* , which we thus named A β^* 56. However, so far our data on the relationship between A β^* 56 and memory impairment are correlative rather than causative. To determine whether A β^* 56 has the ability to disrupt memory function directly, we purified it from the brains of impaired Tg2576 mice, applied it to young rats and assayed its effect using the Morris water maze.

We purified A β^* 56 using immunoaffinity chromatography (with 4G8 antibodies) followed by size-exclusion chromatography. This procedure yielded preparations of A β^* 56 that ran as a single band on silver-stained gels (Fig. 4a), were recognized by 6E10 antibodies (Fig. 4a) and revealed A β as a core component when analysed by liquid chromatography tandem mass spectrometry (Supplementary Fig. 7).

To allow observation of the effects of A β^* 56 on spatial memory for an episode of new learning, our behavioural protocol involved pre-training rats in the Morris water maze, implanting a cannula into a lateral ventricle of the brain, and allowing the animals a two-week recovery period (Supplementary Fig. 6a). We used rats for these experiments because chronic cannulae are better tolerated in rats than in mice. The pre-training allowed for the rapid acquisition of new spatial information in an entirely new environment when similar test procedures were used. Infusing 10 μ l of 0.85 μ M A β^* 56 two hours before testing had no effect on the escape latencies of rats given eight training trials to locate a hidden platform in the new maze environment (Fig. 4b).

After a 24-h retention interval, rats received another injection of A β^* 56 or vehicle, in the same manner as during training, and

underwent a probe trial two hours later during which the platform was absent. Spatial memory was shown by the vehicle-injected rats but not by those treated with A β^* 56. There was a significant interaction of group $\times$ quadrant ($F_{1,18} = 6.288$, $P = 0.022$, analysis of variance (ANOVA)), but no main effects of group or quadrant (Fig. 4d). Rats in the vehicle group spent significantly more time in a target annulus surrounding the platform than in a control annulus in the opposite quadrant ($t_9 = 2.668$, t -test with 9 degrees of freedom; $P = 0.026$). In contrast, rats in the A β^* 56 group spent similar amounts of time in both locations ($P = 0.547$, t -test). In addition, rats in the A β^* 56 group spent significantly less time in the target annulus than those in the vehicle group ($t_{18} = 2.533$, $P = 0.021$, t -test).

Other measures, including the number of crossings (Fig. 4e) and

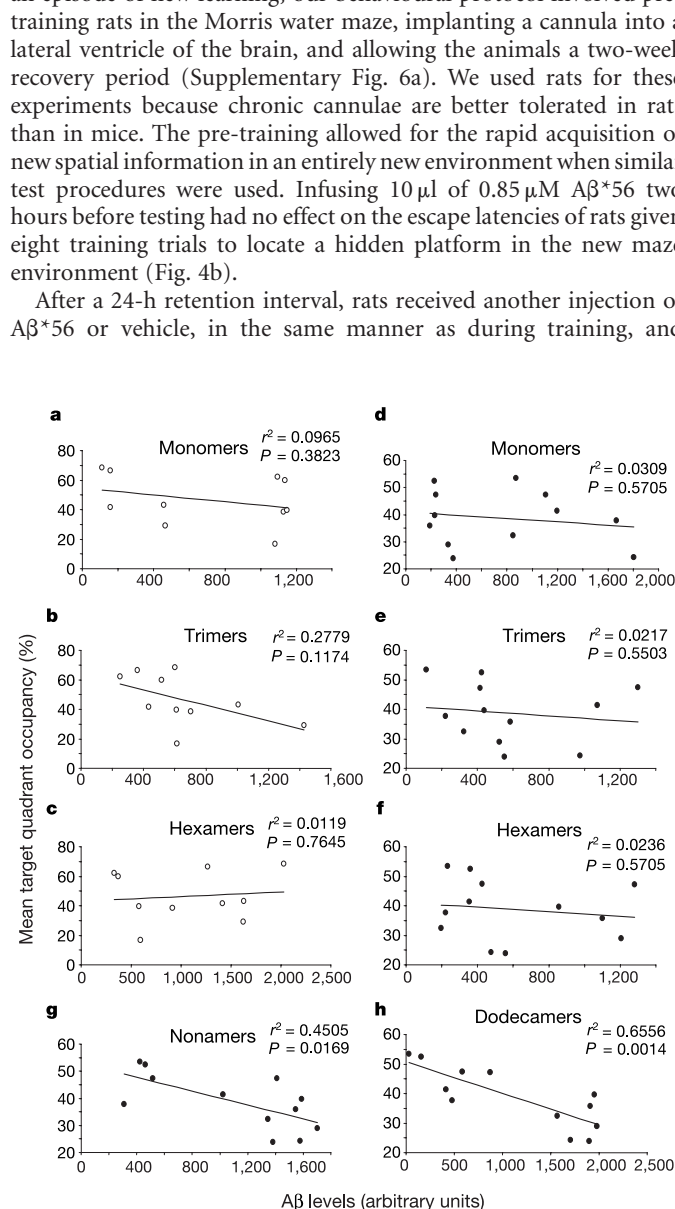


Figure 3 | Levels of the 56-kDa A β assembly show the strongest inverse correlation with spatial memory. **a–f**, Lack of significant correlations between retention of spatial memory and monomeric (4.5 kDa), trimeric (14 kDa) or hexameric (27 kDa) soluble A β species detected in extracellular-enriched fractions of 5-month-old (open circles) and 6-month-old (filled circles) Tg2576^{+/-} mice. **g, h**, Levels of A β assemblies corresponding to nonameric (40 kDa) and dodecameric (56 kDa) species correlate inversely with memory at 6 months of age (ANOVA).

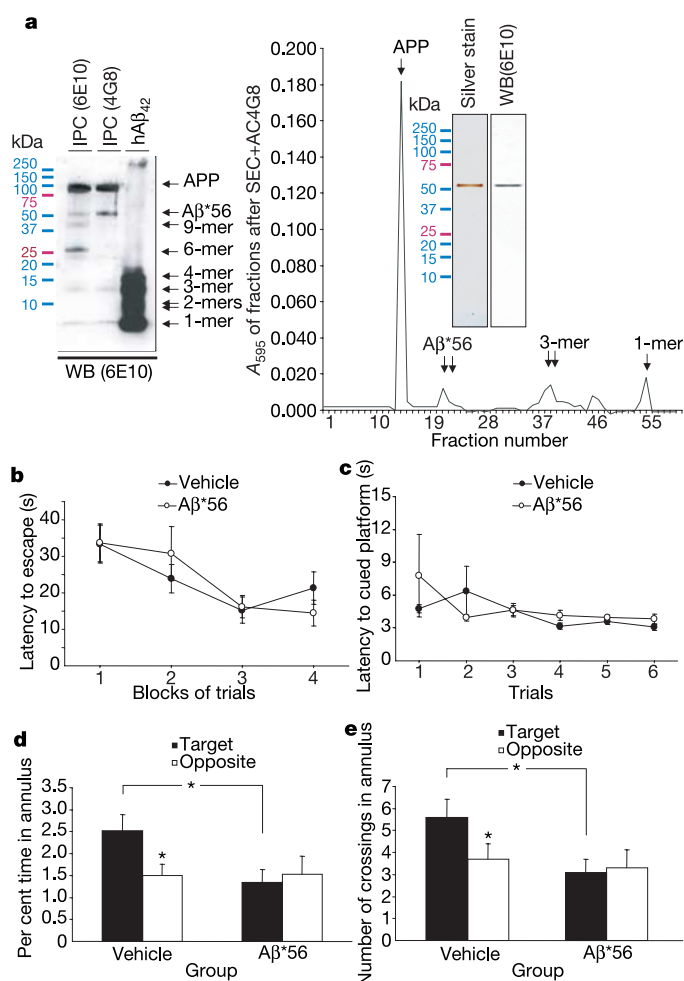


Figure 4 | Purification of A β^* 56 from Tg2576 brain and effects of purified A β^* 56 on memory in young rats. **a**, Purification of total (RIPA-buffer-soluble) A β species using immunoaffinity purification columns (IPC) packed with 200 μ g of 6E10 or 4G8 antibodies (left panel). Right panel shows absorbance at 595 nm (A_{595}) of proteins isolated by affinity chromatography with 4G8 (AC4G8) that were separated by size-exclusion chromatography (SEC) to yield fractions containing A β^* 56. **b**, There is no difference between rats that received A β^* 56 and vehicle (50 mM ammonium acetate, pH 8.5) in terms of latency to locate a hidden platform during training. Data show blocks of two training trials each (mean $\pm$ s.e.m.). **c**, There is no difference between rats (A β^* 56 and vehicle treatment groups) in terms of latency to locate a visible platform. **d, e**, A β^* 56 impairs spatial memory. Rats that received vehicle, but not A β^* 56, injections showed a significant spatial bias for the escape location 24 h after training. Target annulus measures are shown in **d**, and number of annulus crossings in **e** (mean $\pm$ s.e.m.). Two-way ANOVA ($P < 0.05$) followed by t -test (*, $P < 0.05$), $n = 10$ animals per group.

quadrant time, also provided no evidence of spatial bias as an indication of retention in the A β *56 group. Thus, A β *56 impaired long-term memory, but not acquisition of spatial information (Supplementary Discussion). Ten days later, both groups had equally good acquisition and retention in the same two-day protocol when new spatial information was once again provided but no A β *56 was administered, indicating no lasting debilitation as a result of treatment (Supplementary Fig. 6b, c). Notably, the behavioural effects of administering purified A β *56 to rats parallel exactly our previous observations in 6- to 11-month-old Tg2576 mice⁶, which show intact performance during the acquisition phase of the water maze training but impairment in probe trials after an overnight retention interval. Furthermore, in the current test, A β *56 had no effect on escape latencies to a visible platform (Fig. 4c), consistent with the lack of an A β *56 effect on performance of Tg2576 mice in the cued phase of the water maze training (Supplementary Fig. 3). Our data demonstrate that A β *56 impairs memory in healthy, young rats, and strongly support the hypothesis that A β *56 is the principal cause of memory decline in middle-aged Tg2576 mice.

The transient effects of A β *56 application in rats imply that A β *56 impairs memory by inducing transient physiological, rather than permanent neuropathological, alterations of the brain. The accumulation of insoluble aggregating proteins such as tau, huntingtin, prion protein, ataxin and A β occurs in many neurodegenerative disorders. These aggregates often define the disorders neuropathologically, but their relative contribution to disease symptoms compared to other, hypothetical, intermediate protein assemblies is controversial^{21–23}, and the identity of the theoretical intermediates has been elusive. Our finding that A β *56 is responsible for memory loss in plaque-forming Tg2576 mice, and causes memory deficits when injected directly into healthy rats, sets a precedent for identifying other 'star' proteins inducing brain dysfunction—for example, tau* (ref. 22). That A β * is a highly specific form of A β offers the potential for developing precise diagnostic methods to detect its correlate in humans with pre-clinical Alzheimer's disease, presenting the possibility of targeting A β * and aborting the disease before permanent structural changes have developed.

METHODS

Transgenic animals. Tg2576 mice³ were the offspring of mice backcrossed successively to B6SJL/F1 breeders.

Behavioural tests. Spatial reference memory was assessed in Tg2576 mice using a modified version of the Morris water maze⁶. Testing was tailored for Tg2576 transgene-positive and -negative mice in the B6SJL background strain. Tg2576 transgene-positive and -negative mice received visible platform training for three days (eight trials per day), followed by hidden platform training for nine days, with four trials per day (trials 60 s each; inter-trial interval (ITI) 20–40 min). Three probe trials (60 s) were performed 20 h after 12, 24 and 36 training trials (after days 3, 6, and 9 of training), and the mean per cent target quadrant occupancy for the three probe trials was calculated.

Long-Evans male rats, approximately 4 months of age (young adults), were used to test the effects of A β *56 infusions on memory. For pre-training, rats received three trials (90 s; ITI 60 s) per day for eight days, during which they were allowed to locate a camouflaged platform that remained in the same location 2 cm below the water surface in a 1.83-m diameter maze. Every sixth trial was a probe trial (30 s), which served to assess the development of a spatially localized search for the escape platform. Performances during the probe trials were used to generate a learning index as previously described²⁴. Immediately after pre-training, unilateral cannulae were implanted into a lateral ventricle. Coordinates for the cannulae tip were 1.0 mm posterior to bregma, 1.5 mm lateral to midline, and 3.5 mm ventral to the skull surface. Rats were allowed two weeks to recover before being trained to a new platform location in a new water maze environment (with a different physical location and surrounding spatial cues). Ten rats received A β *56 and another ten received vehicle (50 mM ammonium acetate buffer, pH 8.5). Rats in the A β *56 and vehicle groups were matched for body weight and learning index in pre-training. Mean body weights on the training day for rats in A β *56 and vehicle groups were 509.6 $\pm$ 24.43 g (mean $\pm$ s.e.m.) and 525.1 $\pm$ 19.10 g, respectively ($t < 1$). Mean learning indices for rats in the A β *56 and vehicle groups were 202.3 $\pm$ 10.9 and 214.8 $\pm$ 12.4, respectively ($t < 1$). A β *56 was injected at a concentration of 0.85 μ M. The injection volume

was 10 μ l (1.0 μ l min⁻¹, with one additional minute at the end to allow diffusion away from the cannulae). An equal volume of buffer was injected into rats in the vehicle group. Two hours after injection, rats were trained in the water maze. Rats received eight training trials (90 s each; ITI 8 min) to locate a hidden escape platform. The entire training session was completed within 60 min. The retention interval was 24 h. Two hours before the probe trial, A β *56 or vehicle was infused as before training. The percentage of time spent in an annulus that was 1.5 $\times$ the size of the target platform was used to analyse performance in the probe trial. At the end of the probe trial (120 s), rats received six cued training trials to locate a visible escape platform. The location of this platform varied from trial to trial. Ten days after the test with A β *56, both groups were re-tested in the absence of any intraventricular injections before the training session (eight trials) or retention probe trial. The escape platform was in a new location and curtains with new spatial cues surrounded the maze. Both groups improved performance on training and had a similarly accurate spatial search for the correct platform location during the retention test 24 h later (Supplementary Fig. 6b).

Antibodies. The following primary antibodies were used: 6E10 (monoclonal raised against A β _{1–17}) and 4G8 (monoclonal raised against A β _{17–24}) (1:100–10,000 dilution; Signet Laboratories), R1282 (polyclonal raised against A β ; 1:75 dilution)²⁵, R1736 (polyclonal raised against residues 595–611 of APP695; 1:1,000 dilution)²⁶, FCA3542 (polyclonal raised against A β 42)²⁷, APPCter-C17 (polyclonal raised against the carboxy terminus of APP; 1:5,000 dilution)²⁸, anti-Flotillin-2, anti-ERKs, anti-JNK and anti-c-Jun (all at 1:200 dilution; Santa Cruz), anti-tau-5 (1:1,000 dilution; Biosource International), anti-MAP-2 (1:200 dilution; Sigma and Santa Cruz) and anti-actin (1:250 dilution; Sigma), and anti-tPA (anti-tissue plasminogen activator; American Diagnostica). The A11 anti-oligomer antibody (1:5,000 dilution)¹⁸ was detected with a biotinylated anti-rabbit antibody (1:2,000,000 dilution; Vector Laboratories) and ExtrAvidin (1:5,000 dilution; Sigma).

Protein extractions, protein concentration determination, immunoaffinity chromatography, size-exclusion chromatography, silver staining, liquid chromatography tandem mass spectroscopy, western blotting and immunoprecipitation procedures are described in Supplementary Methods.

Received 23 November; accepted 13 December 2005.

1. Craik, F. I. in *Handbook of the Psychology of Aging* (eds Birren, J. E. & Schall, K.) 384–420 (Van Nostrand-Reinhold, New York, 1977).
2. Morrison, J. H. & Hof, P. R. Life and death of neurons in the aging brain. *Science* **278**, 412–419 (1997).
3. Hsiao, K. et al. Correlative memory deficits, A β elevation, and amyloid plaques in transgenic mice. *Science* **274**, 99–102 (1996).
4. Irizarry, M. C., McNamara, M., Fedorchak, K., Hsiao, K. & Hyman, B. T. APPSW transgenic mice develop age-related A β deposits and neuropil abnormalities, but no neuronal loss in CA1. *J. Neuropathol. Exp. Neurol.* **56**, 965–973 (1997).
5. Kawarabayashi, T. et al. Age-dependent changes in brain, CSF, and plasma amyloid β protein in the Tg2576 transgenic mouse model of Alzheimer's disease. *J. Neurosci.* **21**, 372–381 (2001).
6. Westerman, M. A. et al. The relationship between A β and memory in the Tg2576 mouse model of Alzheimer's disease. *J. Neurosci.* **22**, 1858–1867 (2002).
7. Kawas, C. H. et al. Visual memory predicts Alzheimer's disease more than a decade before diagnosis. *Neurology* **60**, 1089–1093 (2003).
8. Small, G. W. et al. Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **97**, 6037–6042 (2000).
9. Bookheimer, S. Y. et al. Patterns of brain activation in people at risk for Alzheimer's disease. *N. Engl. J. Med.* **343**, 450–456 (2000).
10. Albert, M. S. Memory decline: the boundary between aging and age-related disease. *Ann. Neurol.* **51**, 282–284 (2002).
11. Benzing, W. C. et al. Evidence for glial-mediated inflammation in aged APP_{SW} transgenic mice. *Neurobiol. Aging* **20**, 581–589 (1999).
12. Janus, C. et al. A β peptide immunization reduces behavioural impairment and plaques in a model of Alzheimer's disease. *Nature* **408**, 979–982 (2000).
13. Chen, G. et al. A learning deficit related to age and β -amyloid plaques in a mouse model of Alzheimer's disease. *Nature* **408**, 975–979 (2000).
14. Hsia, A. Y. et al. Plaque-independent disruption of neural circuits in Alzheimer's disease mouse models. *Proc. Natl Acad. Sci. USA* **96**, 3228–3233 (1999).
15. Klein, W. L., Krafft, G. A. & Finch, C. E. Targeting small A β oligomers: the solution to an Alzheimer's disease conundrum? *Trends Neurosci.* **24**, 219–224 (2001).
16. Ashe, K. H. Learning and memory in transgenic mice modeling Alzheimer's disease. *Learn. Mem.* **8**, 301–308 (2001).
17. Gordon, J. A. & Warren, J. R. Denaturation of globular proteins. I. The interaction of urea and thiourea with bovine plasma albumin. *J. Biol. Chem.* **243**, 5663–5669 (1968).

18. Kayed, R. *et al.* Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science* **300**, 486–489 (2003).
19. Stern, E. A. *et al.* Cortical synaptic integration *in vivo* is disrupted by amyloid- β plaques. *J. Neurosci.* **24**, 4535–4540 (2004).
20. Billings, L. M., Oddo, S., Green, K. N., McGaugh, J. L. & Laferla, F. M. Intraneuronal A β causes the onset of early Alzheimer's disease-related cognitive deficits in transgenic mice. *Neuron* **45**, 675–688 (2005).
21. Orr, H. T. Neurodegenerative disease: neuron protection agency. *Nature* **431**, 747–748 (2004).
22. SantaCruz, K. *et al.* Tau suppression in a neurodegenerative mouse model improves memory function. *Science* **309**, 476–481 (2005).
23. Duff, K. & Planel, E. Untangling memory deficits. *Nature Med.* **11**, 826–827 (2005).
24. Gallagher, M., Burwell, R. & Burchinal, M. Severity of spatial learning impairment in aging: development of a learning index for performance in the Morris water maze. *Behav. Neurosci.* **107**, 618–626 (1993).
25. Walsh, D. M. *et al.* Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation *in vivo*. *Nature* **416**, 535–539 (2002).
26. Haass, C., Koo, E. H., Mellon, A., Hung, A. Y. & Selkoe, D. J. Targeting of cell-surface β -amyloid precursor protein to lysosomes: alternative processing into amyloid-bearing fragments. *Nature* **357**, 500–503 (1992).
27. Barelli, H. *et al.* Characterization of new polyclonal antibodies specific for 40 and 42 amino acid-long amyloid β peptides: their use to examine the cell biology of presenilins and the immunohistochemistry of sporadic Alzheimer's disease and cerebral amyloid angiopathy cases. *Mol. Med.* **3**, 695–707 (1997).
28. Sergeant, N. *et al.* Progressive decrease of amyloid precursor protein carboxy terminal fragments (APP-CTFs), associated with tau pathology stages, in Alzheimer's disease. *J. Neurochem.* **81**, 663–672 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Younkin, D. Walsh, M. Podlisny, D. Selkoe, J. Cleary and S. Prusiner for critical discussions. We are grateful to D. Cooper-Blacketer, J. McQuail and M. Sherman for technical help, and A. Delacourte and N. Sergeant for providing the APPC17-Cter antiserum. This work was supported by grants from the NIH (to K.H.A., M.G. and A.Y.) and a gift from M. and H. Hobbs to M.G.

Author Contributions S.L. and K.H.A. conceived the project. S.L. planned, performed and analysed the biochemistry experiments, including the protein extractions and purification of A β *56. K.H.A. wrote most of the paper. M.T.K. and M.G. planned, performed and analysed the rat behavioural experiments, and L.K. planned and analysed the mouse behavioural experiments. A.Y. provided the mass spectrometry analysis of purified A β *56, C.G.G. provided biochemistry advice, and C.G.G. and R.K. donated the A11 antiserum.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper at www.nature.com. Correspondence and requests for materials should be addressed to K.H.A. (hsiao005@umn.edu).

LETTERS

Stochastic protein expression in individual cells at the single molecule level

Long Cai^{1*}, Nir Friedman^{1*} & X. Sunney Xie¹

In a living cell, gene expression—the transcription of DNA to messenger RNA followed by translation to protein—occurs stochastically, as a consequence of the low copy number of DNA and mRNA molecules involved^{1–6}. These stochastic events of protein production are difficult to observe directly with measurements on large ensembles of cells owing to lack of synchronization among cells. Measurements so far on single cells lack the sensitivity to resolve individual events of protein production. Here we demonstrate a microfluidic-based assay that allows real-time observation of the expression of β -galactosidase in living *Escherichia coli* cells with single molecule sensitivity. We observe that protein production occurs in bursts, with the number of molecules per burst following an exponential distribution. We show that the two key parameters of protein expression—the burst size and frequency—can be either determined directly from real-time monitoring of protein production or extracted from a measurement of the steady-state copy

number distribution in a population of cells. Application of this assay to probe gene expression in individual budding yeast and mouse embryonic stem cells demonstrates its generality. Many important proteins are expressed at low levels^{7,8}, and are thus inaccessible by current genomic and proteomic techniques. This microfluidic single cell assay opens up possibilities for system-wide characterization of the expression of these low copy number proteins.

β -galactosidase (β -gal) has been the standard reporter for gene expression in both prokaryotic and eukaryotic cells^{9–12}. Only active as a tetramer, a single molecule of β -gal can produce a large number of fluorescent product molecules by hydrolysing a synthetic fluorogenic substrate (Fig. 1b), leading to amplification of the fluorescent signal, as was first demonstrated in 1961 (ref. 13). However, given its potential as a high-sensitivity cellular reporter, β -gal has an Achilles' heel: its fluorescent products are not retained in the cell¹⁰, because efflux pumps on the cell membrane actively and efficiently expel

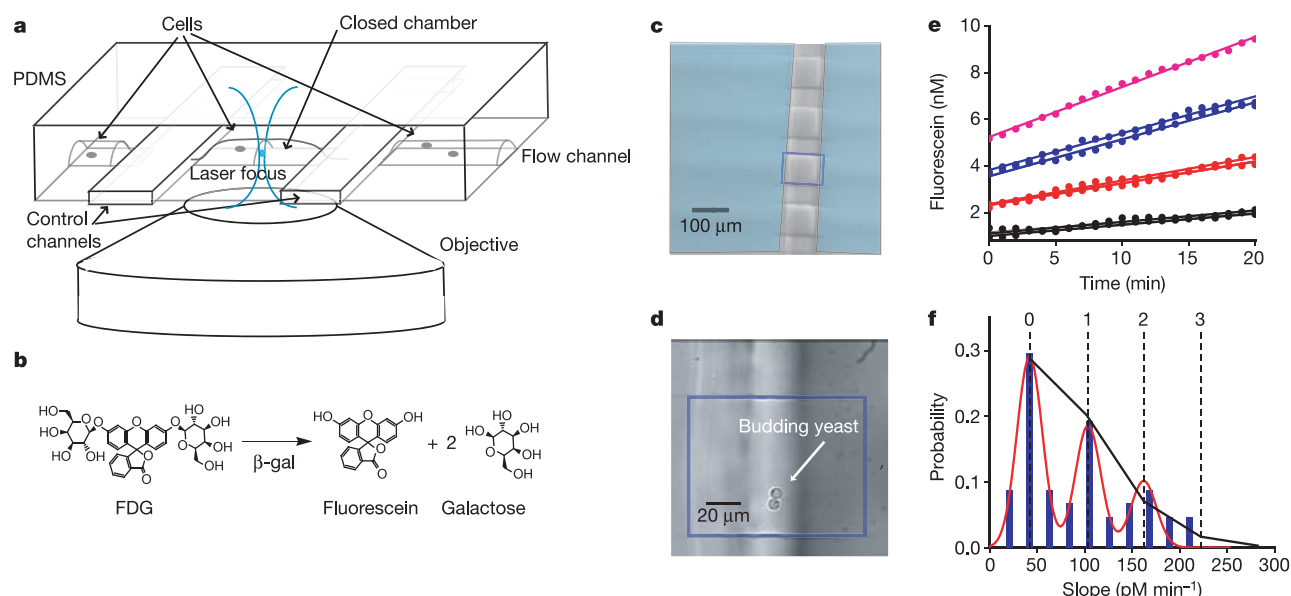


Figure 1 | Single reporter molecule sensitivity in a microfluidic device.

a, Schematic diagram of the microfluidic chamber used for the enzymatic assay. Cells are trapped inside a volume of 100 pL, formed by compression of a flow channel by two control channels. **b**, Enzymatic amplification: hydrolysis of the synthetic substrate FDG by the reporter enzyme β -gal yields a fluorescent product, fluorescein. **c**, Differential interference contrast (DIC) image of a microfluidic chip after actuation of the control channels. The boundaries of one closed chamber are boxed in blue. **d**, DIC image of

budding yeast cells trapped in a chamber. **e**, **f**, FDG hydrolysis by purified β -gal. **e**, Fluorescein concentration increases with time in chambers. The discrete slopes are due to, in decreasing order, 3, 2, 1, 0 (autohydrolysis) β -gal molecules. **f**, A histogram of hydrolysis rates (48 chambers). The red curve is a fit to the data with three Gaussians of equal widths. The peaks are attributed to 0, 1 and 2 enzyme molecules per chamber. The distribution is well-fitted by a Poisson distribution (black line), with an average of 0.7 β -gal molecules per chamber.

¹Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, USA.

*These authors contributed equally to this work.

foreign organic molecules from the cytoplasm¹⁴ (see Supplementary Information). As the fluorescent product molecules are pumped to the surrounding medium and rapidly diffuse away, the advantage of enzymatic amplification is lost.

To circumvent the efflux problem, we trap cells in closed microfluidic chambers, such that the fluorescent product expelled from the cells can accumulate in the small volume of the chambers, recovering the fluorescence signal due to enzymatic amplification. The fast efflux rate and short mixing time of the fluorescent molecules in the miniature chambers guarantee that the fluorescence signal outside the cells accurately reflects the enzymatic activity inside. The microfluidic device is made of a soft polymer, polydimethylsiloxane (PDMS), and consists of a flow layer that contains the cells and a top control layer (Fig. 1a)^{15,16}. Actuation of two adjacent valves in the control layer forms an enclosure of dimensions $100 \times 100 \times 10 \mu\text{m}^3$ (100 pL) in which cells can be trapped and cultured^{17,18} (Fig. 1c, d; see also Supplementary Fig. S1). The microfluidic chip is mounted on an inverted fluorescence microscope and translated by a motorized stage, allowing multiplexing of data acquisition by repeatedly scanning the chambers. Typically, 100 chambers can be scanned within less than 2 min. Fluorescence is excited with a tightly focused laser beam (Fig. 1a) that does not directly illuminate the cell, avoiding cellular autofluorescence and photo-damage to the cell.

We first show the ability to detect single enzyme molecules using this technique by injecting a diluted solution of purified β -gal enzyme and 300 μM of the fluorogenic substrate fluorescein-di- β -D-galactopyranoside (FDG) into the chambers¹⁹. Fluorescence signals from different chambers increase with time, and the slopes give the rates of hydrolysis (Fig. 1e). The distribution of hydrolysis rates measured in the different chambers shows quantized and evenly spaced peaks (Fig. 1f). We attribute these discrete peaks to integer numbers of β -gal molecules. The spacing between the peaks is 60 pM min^{-1} , which gives a calibration for the rate of increase in fluorescein concentration corresponding to one enzyme molecule in a chamber.

Another challenge for using β -gal to monitor gene expression in live cells is that the cell wall acts as a barrier for FDG influx. We quantify this effect in *E. coli* by measuring the hydrolysis rate for live cells compared to cells treated with chloroform, which completely permeabilizes cell membrane (Supplementary Fig. S4a). The ratio of hydrolysis rates between these two cases is defined as the permeability ratio, and is measured to be $R = 13$ at 300 μM FDG. To increase FDG influx, we transformed *E. coli* cells with a plasmid conferring ampicillin resistance and grew the cells in media with β -lactam antibiotics (see Methods). Under these conditions, cell wall synthesis is partially inhibited, making the cells more permeable to FDG, as evident by a lower value of $R = 2 \pm 0.3$ (Supplementary Fig. S4b). In determining the number of enzyme molecules in live cells below, $R = 2$ is used as a correction factor to the *in vitro* calibration (Fig. 1f).

We then monitored gene expression in live *E. coli* cells in real time. β -gal is expressed from the *lacZ* gene on the chromosomal DNA, which is under the control of the *lac* promoter. Cells are grown in glucose-containing medium without inducer to exponential phase; hence the expression level is highly repressed²⁰. We observed abrupt changes in hydrolysis rates in chambers with dividing cells, as shown in Fig. 2a, b. These step-wise increases in the rates indicate the stochastic burst-like expression of new β -gal molecules. We attribute the bursts to stochastic and transient dissociation events of the Lac repressor from the promoter, followed by transcription of mRNA, which is then translated into a few copies of the reporter protein before the mRNA is degraded.

The expression of proteins from a given gene can be characterized by two key parameters: the average frequency of expression bursts per cell cycle, a ; and the average number of protein molecules per burst, b . Under conditions of exponential growth in minimal medium, the

burst frequency for protein expression from the repressed *E. coli lacZ* gene is measured to be $a = 0.11 \pm 0.03$ bursts per cell cycle. The average burst size is measured to be $b = 5 \pm 2$ enzymes, or 20 ± 8 monomers per burst, which is consistent with biochemical estimates of 25–30 β -gal monomers per mRNA^{21,22}.

This real-time assay also allows us to measure the distribution of the number of enzymes produced per burst (Fig. 2c). It can be well fitted with an exponential distribution, $P(n) = C \exp(-n/b)$, where n is the number of β -gal molecules per burst and C is a normalization constant. We attribute this distribution to the fact that the cellular lifetime of the mRNA is exponentially distributed^{21,23}. Previously only theoretically predicted^{1,24}, the exponential $P(n)$ can be accounted for by the competition between mRNA degradation by

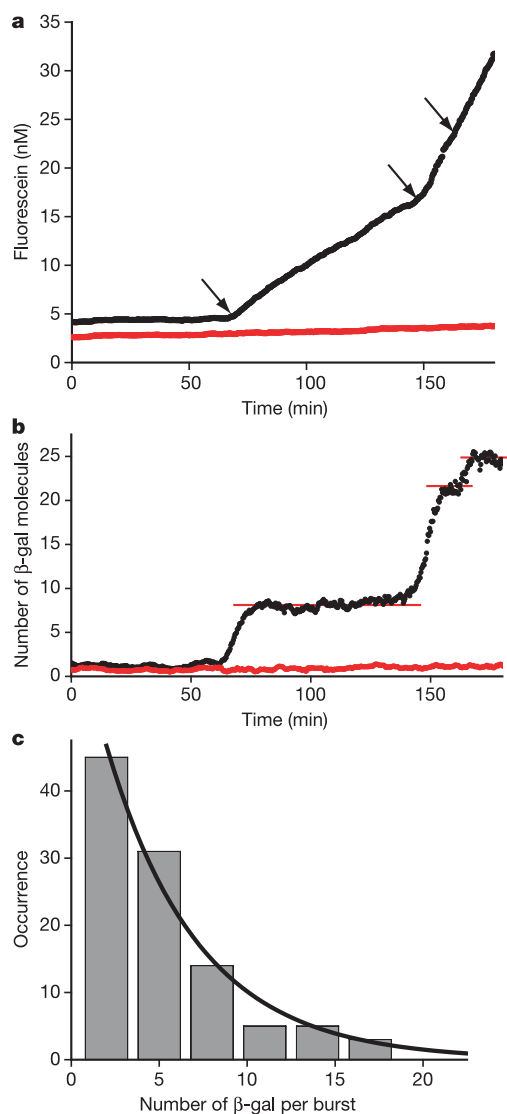


Figure 2 | Quantitative real-time measurement of individual protein expression events in live *E. coli* cells. β -gal is under the control of a repressed *lac* promoter. **a**, Trace of a chamber containing dividing cells shows abrupt changes in hydrolysis rates (arrows on black curve). An empty chamber shows a constant background (red curve). **b**, Discrete jumps in β -gal number are due to burst-like production of proteins. The number of β -gal molecules is calculated by taking the time derivative of the traces in **a** and compensating for fluorescein photobleaching (Supplementary Information). **c**, Histogram of copy number of β -gal molecules per burst. The distribution is well-fitted with an exponential function (black line), with an average of five proteins per burst, and is a consequence of exponential cellular lifetime of the mRNA.

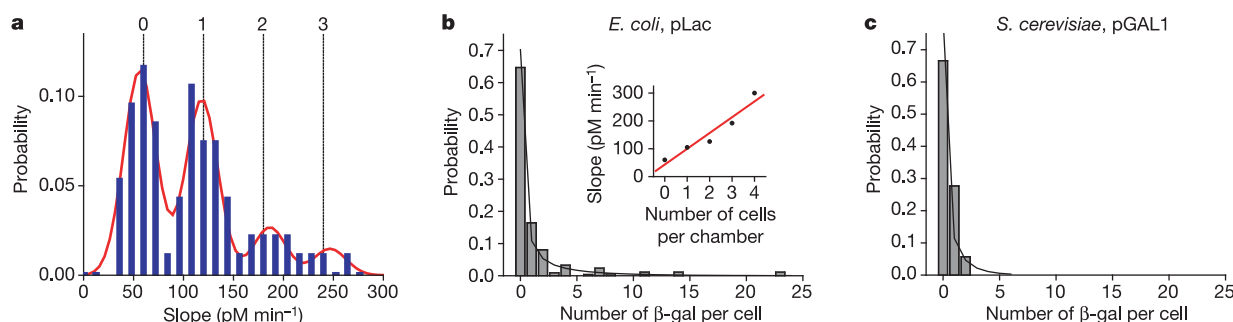


Figure 3 | Steady-state protein copy number distributions in a population of cells. **a**, Histogram of FDG hydrolysis rates from 48 chambers containing chloroform-treated *E. coli* cells shows discrete, evenly spaced peaks, with spacing equal to that measured with purified enzyme (Fig. 1f). **b**, Copy number distribution of β -gal molecules in chambers containing a single *E. coli* cell, under the repressed condition (120 cells). The solid line shows a gamma distribution (equation (1)) with $a = 0.16$ bursts per cell cycle and

$b = 7.8$ β -gal molecules per burst. Inset: average hydrolysis rate grows linearly with the number of cells per chamber with a slope of 0.9 enzymes per cell. **c**, Histogram of the number of β -gal molecules in chloroform-treated yeast cells (116 cells) expressing β -gal from the repressed *GAL1* promoter carried on a low copy number plasmid. The solid line is a gamma distribution with $a = 0.2$ and $b = 1.7$.

RNaseE and translation by ribosome. Our single-molecule experiment demonstrates the ability to provide quantitative real-time information on gene expression in a live cell.

Stochasticity of gene expression is exhibited not only in the temporal evolution of protein production in a single cell, but also in the distribution of protein copy number in a population of cells at a particular time. Such distributions have been measured using flow cytometry^{3,4} or fluorescence microscopy^{2,5,6}, but have been limited to only high expression levels. Our method allows measurements of the reporter's distribution in a cell population with single copy sensitivity.

To do this, we treated *E. coli* cells with chloroform, which completely permeabilizes cell membrane and stops protein production, while keeping existing β -gal molecules active inside the cell (Fig. 3b, inset). The treated cells are immediately injected into the microfluidic device together with FDG and the hydrolysis rate is measured. We first demonstrate the detection of single enzyme molecules inside those chloroform-treated cells. The histogram of hydrolysis rates of all chambers shows quantized peaks (Fig. 3a and Supplementary Fig. S12). The spacing between peaks is in agreement with that measured *in vitro* (Fig. 1f).

We compiled a histogram of the number of β -gal molecules in chambers containing only a single cell (Fig. 3b) for cells cultured under the same repressed condition as in the real-time experiment. The mean of this distribution is $m = 1.2 \pm 0.4$ copies of β -gal molecules per cell, and the standard deviation is $\sigma = 3.3$. The mean agrees with the ensemble measurement of 1 ± 0.5 β -gal per cell, as well as the observation that the average number of β -gal molecules per chamber increases with the number of cells trapped in a chamber with a slope of 0.9 enzymes per cell (Fig. 3b inset).

We created a model to relate the measured steady-state protein copy number distributions to the parameters a and b that are defined above. We used a continuous form of the master equation²⁵ to analytically describe that distribution given the burst dynamics of protein production. We assume that: (1) distribution of protein molecules produced per expression event is exponential; (2) expression events are temporally uncorrelated; and (3) protein molecule copy numbers are halved at cell division. At steady-state, protein production is balanced by protein dilution due to cell growth and division. We found that the steady-state probability distribution of protein number per cell, $p(x)$, is a gamma distribution, which is uniquely determined by the parameters a and b (Supplementary Information):

$$p(x) = \frac{x^{a-1} e^{-x/b}}{b^a \Gamma(a)} \quad (1)$$

where Γ denotes the gamma function. The parameters a and b are

related to the mean, m , and standard deviation, σ , by $a = m^2/\sigma^2$ and $b = \sigma^2/m$. Figure 3b shows that the data fits well with equation (1), with $b = 7.8 \pm 2.6$ reporters per burst and $a = 0.16 \pm 0.05$ bursts per cell cycle, which are consistent with the real-time data above within error bars. This demonstrates that the information about the underlying dynamic process, namely the values of a and b , is preserved in both the temporal fluctuations in a single cell and the instantaneous variability within a population.

Notably, the profile of $p(x)$ with $a < 1$ that we measured for the repressed *lac* operon is distinct from $p(x)$ with $a > 1$ that was measured previously for the induced *lac* operon². These correspond to two different regimes of equation (1), schematically illustrated in Fig. 4. For $a < 1$, $p(x)$ peaks at zero and a large fraction of cells contains no β -gal regardless of the burst size b (Fig. 4a); however, for

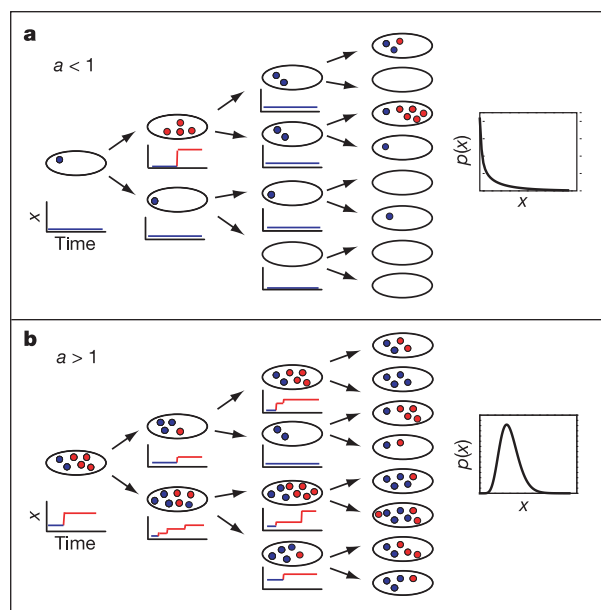


Figure 4 | Two regimes of stochasticity in protein expression. Protein molecules are produced in bursts (red), in addition to existing molecules (blue), and diluted upon cell division, leading to a steady-state distribution, $p(x)$, of protein copy number, x , in the cell population. **a**, With $a < 1$, $p(x)$ peaks at $x = 0$ and a large fraction of cells do not contain a single copy of the protein regardless of the burst size, b . The repressed *lac* operon is well described by this regime. **b**, With $a > 1$, most cells contain the protein at some level. Both regimes can be described by the gamma distribution (equation (1)).

$a > 1$, $p(x)$ peaks at a non-zero x and most cells contain β -gal (Fig. 4b).

The profile of $p(x)$ for $a < 1$ may be significant for the response of a population of *E. coli* cells to an increase in lactose concentration, which is transported into the cell by Lac permease. Because β -gal and Lac permease are expressed from the same operon, the large sub-population of cells containing no β -gal would generally have no Lac permease. These cells cannot respond to a rise of lactose concentration in the environment, and thus are distinct from cells containing permease, which can respond and induce further expression of the *lac* operon by way of positive feedback. Thus, $a < 1$ provides a mechanism for 'all-or-none' induction responses that cells exhibit under suboptimal inducer concentrations^{11,26}.

This microfluidic assay can be applied to other cell types expressing β -gal as a reporter. We demonstrate applicability by measuring low-level expression from repressed promoters in yeast and mammalian cells. Figure 3c shows the distribution of reporter number in a population of chloroform-treated *Saccharomyces cerevisiae* (yKT32) cells expressing β -gal from the repressed *GALI* promoter. Here, the promoter-reporter fusion is carried on a low copy number plasmid. Cells were grown in the presence of glucose, such that the promoter is expected to be highly repressed. The average expression level is measured to be 0.4 ± 0.07 reporters per cell, with $\sigma = 0.6$. Using the above model (equation (1)) we can extract a and b from the measured mean and σ : $a = 0.2 \pm 0.08$ bursts per cell cycle and $b = 1.7 \pm 0.4$ reporters per burst. This is consistent with our real-time experiments with permeable mutant yeast strain (*pdr1⁻ pdr3⁻*) in which we do not observe abrupt large jumps as seen with *E. coli*. The low burst size in this case may reflect the non-optimal codon usage of the bacterial reporter gene in yeast. Finally, we measured β -gal expression also from the ROSA26 promoter in embryonic mouse stem cells (ES 54A). The average expression level was 425 ± 80 enzymes per cell with $\sigma = 375$ (Supplementary Information).

The single cell version of the β -gal assay⁹ presented here exhibits single molecule sensitivity for single prokaryotic and eukaryotic cells. This method, together with other newly developed single cell and single molecule techniques at the mRNA^{27,28} and protein²⁹ levels, opens up possibilities for studying low-level gene expression in single cells of diverse cell types (for example, transcription factors involved in gene regulation, or the leakiness of supposedly silenced promoters in differentiated cells). The ease of scaling up the microfluidic assay will enable characterization of gene expression on a whole-genome scale, especially in the unexplored regime of low expression levels.

METHODS

Cell strains and growth conditions. *E. coli* K12 (MG1655) cells are transformed with a medium copy number plasmid, pBR293.3, which confers ampicillin resistance. Cells were cultured in M9 media supplemented with amino acids and vitamins. Experiments were performed at 30 °C. Under these conditions cells grow with a cell cycle time of about 145 min (Supplementary Fig. S1). *S. cerevisiae* cells (yKT0032) express β -gal from the *GALI* promoter, carried on a low-copy 2 μ m plasmid (pYES2-lacZ). Cells were cultured in SD glucose –ura medium (BD Bioscience) at 30 °C. Mouse embryonic stem cells ES54A (Smo +/–, Rosa26 lacZ⁺) and ES17 (lacZ[–]) from a 129SVJ background were cultured in a defined medium to prevent differentiation. Single cell suspension is derived by adding trypsin to the cell culture and resuspending.

Microfluidic chip fabrication. The master mould for the bottom channels is made by spin-coating positive photoresist (Shipley SPR 220-7, MicroChem) onto a silicon wafer (University Wafer) to a height of 7 μ m. After patterning with high-resolution transparency mask (PageWorks), the features are rounded by heating^{15,16}. The control layer master is patterned with negative photoresist at 40 μ m height. PDMS (10:1 Dow Corning Sylgard 184 A:B, Ellsworth Adhesives) is spun onto the bottom master at 2,000 r.p.m. for 1 min and partially cured. Control layer PDMS mould is also partially cured and baked at the thickness of 5 mm. The two PDMS pieces are aligned and cured overnight. The two-layer device is cut out, plasma oxidized for 3 min and bonded permanently to a coverslip.

Fluorescence detection. Fluorescence was excited by a tightly focused beam of a 488-nm Ar-ion laser (Melles Griot), with a power of 27 μ W. A high NA objective

lens (Nikon 60 $\times$ Plan Apo NA = 1.4) was used in an inverted fluorescence microscope (Nikon) for excitation and collection of the fluorescence signal. A CCD camera (Andor Technology) with on-chip amplification and a large dynamic range was used for detection. Filters used were: excitation, 488 nm narrow bandpass; emission, 535 nm/50 bandpass; dichroic, 495 nm high pass (Chroma Corp). Measurement of signals from multiple chambers was multiplexed by scanning the microfluidic chip using a motorized stage, ASI MS2000 XY (ASI).

Data analysis. Analysis of the fluorescent signal was performed using Matlab (Mathworks). Signal was averaged over the 100 brightest pixels from each image, and background was subtracted. The measured fluorescence time traces $F(t)$ were corrected for photobleaching and time derivative was taken to obtain the number of enzymes in a chamber as a function of time. A median filter of length 11 was used to smooth $F(t)$. With a sampling rate of once every 20 s per chamber, the effective time resolution is about 4 min. The error in calculating the parameters a and b from the copy number distribution was estimated with bootstrap analysis³⁰. The original data set was re-sampled 1,000 times. The a and b parameters were calculated for each of those re-sampled data sets, from which the error was estimated. Additional details are provided in the Supplementary Information.

Received 12 September 2005; accepted 23 January 2006.

- McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94**, 814–819 (1997).
- Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
- Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31**, 69–73 (2002).
- Blake, W. J., Kærn, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633–637 (2003).
- Raser, J. M. & O'Shea, E. K. Control of stochasticity in eukaryotic gene expression. *Science* **304**, 1811–1814 (2004).
- Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962–1965 (2005).
- Guptasarma, P. Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of *Escherichia coli*? *Bioessays* **17**, 987–997 (1995).
- Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
- Miller, J. H. *Experiments in Molecular Genetics* (Cold Spring Harbor Laboratory, New York, 1972).
- Russo-Marie, F., Roederer, M., Sager, B., Herzenberg, L. A. & Kaiser, D. β -galactosidase activity in single differentiating bacterial cells. *Proc. Natl Acad. Sci. USA* **90**, 8194–8198 (1993).
- Novick, A. & Weiner, M. Enzyme induction as an all-or-none phenomenon. *Proc. Natl Acad. Sci. USA* **43**, 553–566 (1957).
- Nolan, G. P., Fiering, S., Nicolas, J. F. & Herzenberg, L. A. Fluorescence-activated cell analysis and sorting of viable mammalian cells based on β -D-galactosidase activity after transduction of *Escherichia coli* lacZ. *Proc. Natl Acad. Sci. USA* **85**, 2603–2607 (1988).
- Rotman, B. Measurement of activity of single molecules of β -D-galactosidase. *Proc. Natl Acad. Sci. USA* **47**, 1981–1991 (1961).
- Nikaido, H. Prevention of drug access to bacterial targets: permeability barriers and active efflux. *Science* **264**, 382–388 (1994).
- Unger, M. et al. Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science* **288**, 113–116 (2000).
- Wu, H., Wheeler, A. & Zare, R. N. Chemical cytometry on a picoliter-scale integrated microfluidic chip. *Proc. Natl Acad. Sci. USA* **101**, 12809–12813 (2004).
- McDonald, J. C. et al. Fabrication of microfluidic systems in poly(dimethylsiloxane). *Electrophoresis* **21**, 27–40 (2000).
- Balagadde, F. K., You, L., Hansen, C. L., Arnold, F. H. & Quake, S. R. Long-term monitoring of bacteria undergoing programmed population control in a microchemostat. *Science* **309**, 137–140 (2005).
- Rondelez, Y. et al. Microfabricated arrays of femtoliter chambers allow single molecule enzymology. *Nature Biotechnol.* **23**, 361–365 (2005).
- Monod, J., Pappenheimer, A. M. Jr & Cohen-Bazire, G. The kinetics of the biosynthesis of β -galactosidase in *Escherichia coli* as a function of growth. *Biochim. Biophys. Acta* **9**, 648–660 (1952).
- Kennell, D. & Riezman, H. Transcription and translation initiation frequencies of the *Escherichia coli* lac operon. *J. Mol. Biol.* **114**, 1–21 (1977).
- Sorensen, M. & Pedersen, S. Absolute in vivo translation rates of individual codons in *Escherichia coli*. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J. Mol. Biol.* **222**, 265–280 (1991).
- Berg, O. G. A model for the statistical fluctuations of protein number in a microbial population. *J. Theor. Biol.* **71**, 587–603 (1975).
- Yarchuk, O., Jacques, N., Guillerez, J., Dreyfus, M. Interdependence of translation, transcription and mRNA degradation in the *lacZ* gene. *J. Mol. Biol.* **226**, 581–596 (1992).

25. Paulsson, J., Berg, O. G. & Ehrenberg, M. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc. Natl Acad. Sci. USA* **97**, 7148–7153 (2000).
26. Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & Van Oudenaarden, A. Multistability in the lactose utilization network of *Escherichia coli*. *Nature* **427**, 737–740 (2004).
27. Levsky, J. M., Shenoy, S. M., Pezo, R. C. & Singer, R. H. Single-cell gene expression profiling. *Science* **297**, 836–840 (2002).
28. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–1036 (2005).
29. Yu, J. *et al.* Probing gene expression in live cells, one protein molecule at a time. *Science* **311**, 1600–1603 (2006).
30. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge J. Xiao, K. Gudiksen and J. Markson for early involvement in the project, and J. Yu, P. Choi and J. Elf for discussions and careful reading of the manuscript. We are grateful to Q. Xu, H. Wu, members of G. M. Whitesides and R. N. Zare groups, and the Harvard Center for Nanoscale Systems for help with microfluidic fabrication. We thank K. Thorn for yeast strains, and A. and J. McMahon for stem cell strains. This work was funded by Department of Energy, Office of Biological and Environmental Research, Genomics: GTL Program, and in part by National Institute of Health (NIH) Director's Pioneer Award to X.S.X., and an NIH grant. L.C. is supported by a National Science Foundation Graduate Research Fellowship and N.F. by a Human Frontiers Science Program Long Term Fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to X.S.X. (xie@chemistry.harvard.edu).

The PerR transcription factor senses H_2O_2 by metal-catalysed histidine oxidation

Jin-Won Lee¹ & John D. Helmann¹

The sensing of reactive oxygen species is essential for cellular responses to oxidative stress¹. The sensing of peroxides is typically mediated by redox-active cysteines in sensors such as the bacterial OxyR, OhrR, and Hsp33 proteins^{2,3}. *Bacillus subtilis* PerR is the prototype for a widespread family of metal-dependent peroxide sensors that regulate inducible peroxide-defence genes⁴. Here we show that PerR senses peroxides by metal-catalysed oxidation. PerR contains two metal-binding sites: a structural Zn^{2+} site and a regulatory divalent metal ion site that preferentially binds Fe^{2+} or Mn^{2+} (ref. 5). Protein oxidation, catalysed by a bound ferrous ion, leads to the rapid and direct incorporation of one oxygen atom into histidine 37 (H37) or H91, two of the residues that coordinate the bound Fe^{2+} . This mechanism accounts for the ability of PerR to sense low levels of hydrogen peroxide *in vivo*. The reduction of hydrogen peroxide by metal ions to generate highly reactive hydroxyl radicals underlies the genotoxic effects of peroxides¹, and has been shown to contribute to enzyme inactivation, but has not previously been shown to provide a regulatory mechanism for peroxide sensing.

Bacillus subtilis PerR is related to the ferric-uptake repressor (Fur) family of proteins⁴. Under most growth conditions the PerR dimer contains two metal ions per subunit (PerR:Zn,Fe) (ref. 5) and PerR-regulated genes are rapidly induced upon exposure to low levels ($<10 \mu\text{M}$) of hydrogen peroxide (H_2O_2)⁶. In contrast, in iron-limited medium supplemented with Mn^{2+} the resulting PerR:Zn,Mn form of the repressor is relatively insensitive to H_2O_2 (ref. 7). Derepression of the PerR regulon (which includes catalase, alkyl-hydroperoxide reductase, and the Dps-like DNA-binding protein MrgA) acts to detoxify peroxides and minimize the damage from intracellular Fe-catalysed redox reactions. By analogy with other H_2O_2 sensors (OxyR, RsrA, Hsp33, yeast Yap1, and eukaryotic protein tyrosine phosphatases)^{2,3,8}, it has been hypothesized that cysteine residues are involved in H_2O_2 sensing by PerR⁵ and its orthologues, CatR⁹ and FurS¹⁰. While cysteine oxidation of these proteins has been demonstrated *in vitro*, the concentrations of H_2O_2 used in these studies were far in excess of those that elicit biological responses.

To provide a structural framework for the investigation of possible peroxide-sensing mechanisms, we modelled the PerR protomer (Fig. 1a) based on the *Pseudomonas aeruginosa* Fur structure (FurPA)¹¹. In PerR, four cysteine residues (C96, C99, C136 and C139) cluster to form a high-affinity Zn^{2+} -binding site (Cys₄Zn) not found in FurPA. Five other residues (H37, D85, H91, H93 and D104) are candidate ligands for the regulatory metal ion, Fe^{2+} or Mn^{2+} . These assignments are supported by site-directed mutagenesis studies: all nine residues proposed to be metal ligands are essential for repressor function *in vivo* (Fig. 1b and Supplementary Fig. S1).

To define the mechanism of peroxide sensing we purified PerR together with its structural Zn^{2+} (PerR:Zn) and reconstituted PerR,

in both the Fe^{2+} and Mn^{2+} cofactored forms. When purified in the presence of high concentrations of EDTA (Fig. 2a, trace 1), PerR contained nearly stoichiometric amounts of Zn^{2+} and was highly active (see below). Protein purified without EDTA (trace 4), or generated by exposure of purified PerR:Zn to Fe^{2+} (trace 2) was oxidized, as judged

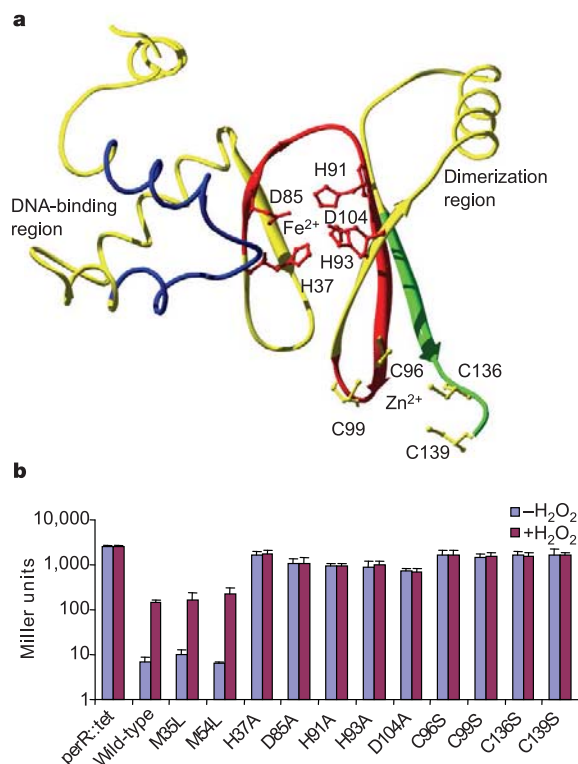


Figure 1 | PerR contains two metal-binding sites. **a**, A monomer of PerR was modelled on the *Pseudomonas aeruginosa* Fur (FurPA) structure¹¹ using SwissModel²⁷. Candidate amino acid ligands for Fe^{2+} (red) and Zn^{2+} (yellow) are conserved in PerR family members (see Supplementary Fig. S1a). The sites of H_2O_2 -mediated oxidation (H37 and H91) are in tryptic peptides 1 (T1, shown in blue) and 2 (T2, red) and are positioned to bind Fe^{2+} (corresponding to FurPA metal site 2). The four cysteine residues are in peptides T2 (red) and T3 (green). **b**, Mutational analyses indicate that the Zn^{2+} - and Fe^{2+} -binding amino acids are essential for repressor function (see also Supplementary Fig. S1). Repression was measured using an *mrgA-cat-lacZ* fusion reporter construct and are expressed as Miller units of β -galactosidase activity (values represent the mean $\pm$ s.d. from three separate experiments). Activity was measured in a *perR::tet* null mutant (left bars) or after complementation of a null mutant with various *perR* alleles as indicated.

¹Department of Microbiology, Cornell University, Ithaca, New York 14853, USA.

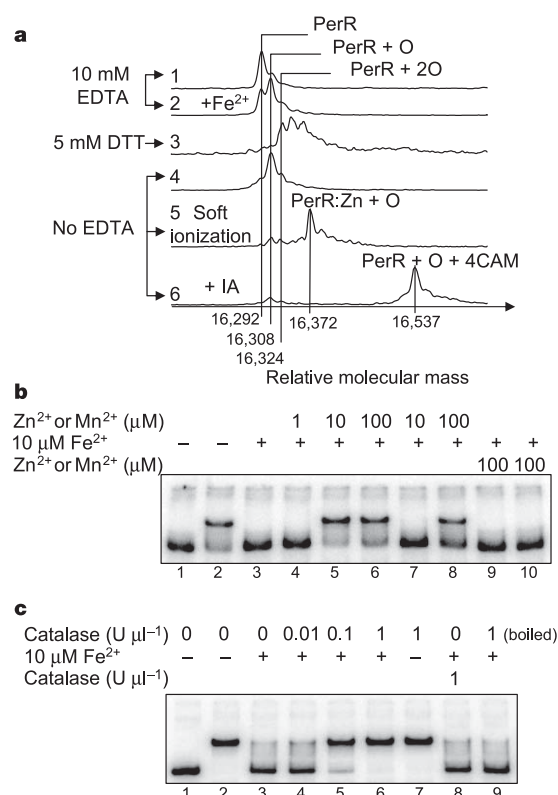


Figure 2 | Metal-catalysed oxidation of PerR. **a**, Analysis of protein oxidation by ESI-MS. PerR purified in the presence of 10 mM EDTA is mostly unoxidized (trace 1) and is >60% active. PerR purified without EDTA (trace 4) or aerobically exposed to excess Fe²⁺ for 10 min after purification (trace 2) contains one (O) or two (2O) additional oxygens and is largely inactive. When purified in buffers lacking EDTA and containing 5 mM dithiothreitol (DTT) (trace 3), PerR is oxidized at multiple sites (see Supplementary Fig. S8). Oxidized PerR (trace 4) contains Zn²⁺ (trace 5) and all four cysteines are reduced (trace 6). **b**, Oxidation of PerR:Zn by Fe²⁺ monitored using EMSA with *mrgA* promoter region DNA. DNA mobility (lane 1) is reduced by the addition of PerR:Zn, which binds Mn²⁺ present in the binding buffer and gel to form the active PerR:Zn,Mn complex (lane 2). 1 μM PerR:Zn was exposed aerobically to 10 μM Fe²⁺ as indicated and diluted (20 ×) prior to analysis by EMSA. Inactivation of PerR:Zn by Fe²⁺ (lane 3) can be prevented by Zn²⁺ (lanes 4–6) or Mn²⁺ (lanes 7 and 8), but not when added after Fe²⁺ (lanes 9 and 10). All lanes contain <1 nM DNA and 25 nM PerR dimer except lane 1, which contains DNA only. **c**, EMSA shows that the aerobic inactivation of PerR:Zn by Fe²⁺ (lane 3) is blocked by catalase (lanes 4–6), but only when present and active during treatment (lanes 8–9). Lane 1 contains *mrgA* promoter region DNA only. 1 μM PerR:Zn was exposed aerobically to 10 μM Fe²⁺ as indicated and diluted (20 ×) prior to analysis by EMSA.

by an increase in mass corresponding to one to two oxygen atoms. These results suggest that protein oxidation during purification can be reduced by the removal of loosely associated metal ions. Protein purified in the absence of EDTA was largely oxidized and inactive, but still contained a full complement of Zn²⁺ (trace 5) and all four cysteine residues were fully modified by iodoacetamide, indicating that they were reduced (trace 6). Thus, oxidative inactivation of PerR does not appear to affect the structural Cys₄Zn site.

The high sensitivity of PerR:Zn,Fe to oxidation was apparent during initial attempts to reconstitute the protein aerobically. We monitored DNA-binding by PerR:Zn,Mn under aerobic conditions using electrophoretic mobility shift assays (EMSA) with buffers containing Mn²⁺ (Fig. 2b, lane 2). Exposure of PerR:Zn to Fe²⁺ under aerobic conditions inactivated the protein (Fig. 2b, lane 3). This inactivation was site-specific, and could be blocked by Zn²⁺ or Mn²⁺ (Fig. 2b, lanes 4–8), which compete for the Fe²⁺-binding site,

or by catalase (Fig. 2c). Under these conditions, micromolar levels of H₂O₂ were generated by auto-oxidation of Fe²⁺ and by the subsequent dismutation of the initially formed superoxide anion. Thus, PerR is highly sensitive to metal-catalysed oxidation (MCO) by bound Fe²⁺ and H₂O₂, which leads to oxygen incorporation into the protein.

PerR was highly sensitive to H₂O₂-mediated inactivation *in vivo* in medium containing iron, but not in low-iron medium supplemented with manganese (Fig. 3a), consistent with previous findings^{7,12}. To compare directly the H₂O₂ sensitivity of PerR:Zn,Mn and PerR:Zn,Fe, both proteins were reconstituted anaerobically. In a fluorescence anisotropy (FA) DNA-binding assay, PerR:Zn,Mn was insensitive to H₂O₂ whereas PerR:Zn,Fe rapidly lost activity in response to H₂O₂ (Fig. 3b). From these studies, we estimate a second-order rate constant of ~10⁵ M⁻¹ s⁻¹ for inactivation of PerR:Zn,Fe by H₂O₂, which is comparable to the reported rate of oxidation of OxyR by H₂O₂ (ref. 13). DNA-binding by oxidized PerR:Zn,Fe could not be restored by the addition of Mn²⁺ (Fig. 3b, lane 5). Similarly, using EMSA we showed that PerR:Zn,Fe was ~95% inactivated by two molar equivalents (100 μM) of H₂O₂ within 1 min, whereas PerR:Zn,Mn was resistant to 10 mM H₂O₂ (Fig. 3c). We note that the fractional inactivation of reconstituted PerR:Zn,Fe by H₂O₂ was greater than the observed oxidation of Fe²⁺ to Fe³⁺ (Fig. 3d), consistent with an oxidation event that can regenerate reduced ferrous ion (see below). We conclude that PerR:Zn,Fe is >10⁴-fold more sensitive than PerR:Zn,Mn to H₂O₂ inactivation.

We used mass spectrometry to identify sites of PerR oxidation *in vitro* and *in vivo*. When anaerobically reconstituted PerR:Zn,Fe was treated with isotopically enriched H₂¹⁸O₂, oxygen was incorporated into two tryptic peptides (T1 and T2*, Fig. 4a) predicted to contain residues that coordinate Fe²⁺ (Fig. 1). Both peptides that contain cysteine residues (T2* and T3*) were fully modified by iodoacetamide (indicating the presence of reduced cysteine residues) even after H₂O₂ treatment, as is consistent with the refractory nature of the Cys₄Zn site to oxidation. No oxidation was observed with reconstituted PerR:Zn,Mn under identical conditions (Supplementary Fig. S3b).

To investigate H₂O₂ sensing *in vivo*, we expressed fully functional FLAG epitope-tagged PerR (PerR-FLAG) in *B. subtilis*. We purified PerR-FLAG from cells with or without H₂O₂ (or H₂¹⁸O₂) treatment and, as observed *in vitro*, the T1 and T2* peptides had each incorporated oxygen (Fig. 4b). As noted *in vitro*, there was no evidence of cysteine oxidation (Supplementary Figs S2 and S3e). The efficient incorporation of ¹⁸O into PerR (Fig. 4a and b) demonstrates that the hydroxyl radical (generated by the reduction of H₂O₂ by bound Fe²⁺) reacts directly with the protein. Significantly, oxidation was greatly reduced in cells grown in iron-limited minimal medium supplemented with Mn²⁺ (Supplementary Fig. S3d), which is consistent with the reduced reactivity of PerR under these growth conditions (Fig. 3a). Moreover, oxidation of H37 *in vivo* was eliminated in mutants containing an alanine (A) at amino acid position 91 instead of a histidine (H91A) and H91 oxidation was eliminated in the H37A mutants. These data are consistent with the hypothesis that oxidation requires site-specifically bound Fe²⁺ (Supplementary Fig. S4). As befitting its role as the primary regulator of peroxide defences, PerR is much more sensitive to oxidation than is Fur, its Fe²⁺-sensing paralogue: H₂O₂ does not efficiently induce the Fur regulon *in vivo*⁶ nor does it oxidize the corresponding histidine residues (Supplementary Fig. S5).

Sequencing of the oxidized tryptic peptides by electrospray ionization tandem mass spectrometry (ESI-MS) reveals that *in vivo* ¹⁸O is incorporated into H37 (on T1) or H91 (on T2), two residues predicted to coordinate Fe²⁺ (Fig. 4c and see Supplementary Figs S6 and S7 for corresponding *in vitro* data). Oxidation of methionine 35 (M35) also occurs at a low level *in vitro* and thereby contributes to both the T1 + O (<20%) and T1 + 2O ions (Supplementary Figs S6 and S7). However, mutagenesis of this residue

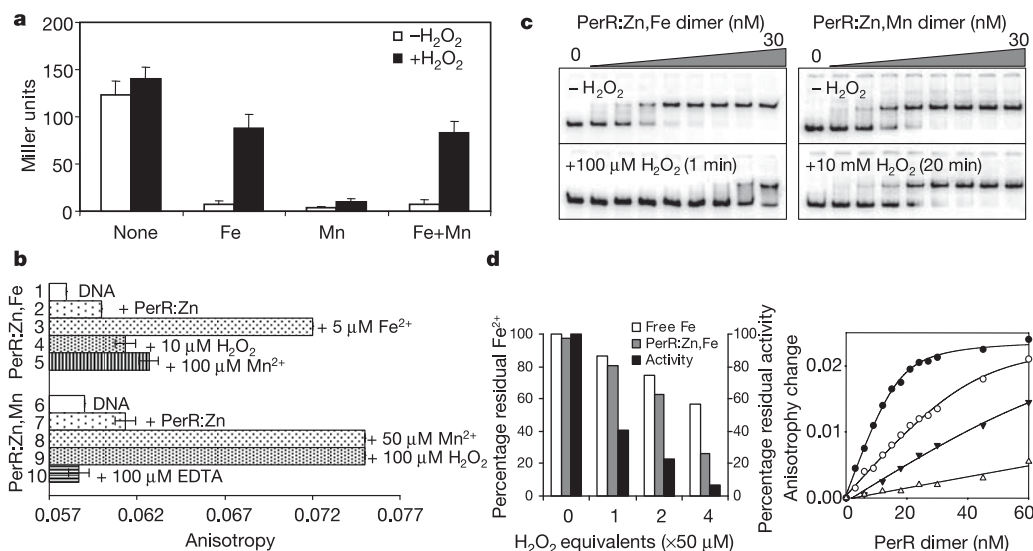


Figure 3 | Oxidative inactivation of PerR:Zn,Fe and PerR:Zn,Mn. **a**, PerR inactivation by H_2O_2 *in vivo* requires iron and is blocked by high manganese. Repression is measured using an *mrgA-cat-lacZ* fusion reporter construct and is expressed as Miller units of β -galactosidase activity (values represent the mean $\pm$ s.d. from three separate experiments). Cells were grown in minimal medium containing 100 μM Fe^{2+} and/or Mn^{2+} as indicated. **b**, Purified PerR:Zn (Fig. 2a, trace 1) was reconstituted anaerobically with Fe^{2+} (lane 3) or Mn^{2+} (lane 8) and DNA binding monitored by FA. All lanes contain 100 nM of fluorescein-labelled *mrgA* Per box DNA and 160 nM PerR:Zn, with the exception of lanes 1 and 6, which contain DNA only. PerR:Zn,Fe was rapidly inactivated (<20 s) by 10 μM H_2O_2 (lane 4) and was not reactivated by Mn^{2+} (lane 5). PerR:Zn,Mn was

insensitive to 100 μM H_2O_2 for 10 min (lane 9). Error bars represent mean $\pm$ s.d. of three readings. **c**, Treatment of 50 μM PerR:Zn,Fe with 100 μM H_2O_2 for 1 min led to $>95\%$ inactivation (K_d decreased from ~ 1 to ~ 30 nM) as judged by EMSA. In contrast, PerR:Zn,Mn retained $>50\%$ activity after exposure to 10 mM H_2O_2 for 20 min. **d**, Fe^{2+} or PerR:Zn,Fe were incubated for 20 s with H_2O_2 and the residual Fe^{2+} content was measured (PerR:Zn stimulates Fe^{2+} oxidation relative to Fe^{2+} alone). Residual DNA-binding activity (relative to the control; left panel) was determined using FA (right panel): the calculated percentage activity was 58, 24, 13 or 4 after treatment with 0 (filled circles), 1 (open circles), 2 (filled triangles), or 4 (open triangles) equivalents of H_2O_2 (corresponding to 0, 50, 100 and 200 μM H_2O_2 , respectively).

(M35L) does not impair peroxide sensing by PerR (Fig. 1b). These observations confirm that PerR:Zn,Fe senses H_2O_2 by direct MCO of H37 or H91.

By analogy to MCO reactions with histidine-containing enzymes and model peptides, we suggest that oxidation of PerR leads to the formation of 2-oxo-histidine residues^{14–17} (Fig. 4d). The proposed pathway of oxygen incorporation¹⁶ results in an intermediate 2-oxo-histidyl radical that could regenerate Fe^{2+} . Regeneration of Fe^{2+} may account for the disparity between the fractional inactivation of PerR and the oxidation of Fe^{2+} to Fe^{3+} noted *in vitro* (Fig. 3d).

Inactivation of PerR:Zn,Fe by H_2O_2 could be explained by oxidation and loss of the activating Fe^{2+} , protein oxidation, or both. Because excess Mn^{2+} is unable to restore DNA-binding activity to oxidized PerR (Fig. 3b and c), we favour a model in which protein oxidation is sufficient for induction. In principle, the regeneration of reduced iron (Fig. 4d) could allow additional cycles of MCO, thereby leading to multiply oxidized proteins (as observed *in vitro*; Fig. 2a and Supplementary Fig. S8). However, this process does not proceed efficiently *in vivo* where $\sim 50\%$ oxidation of both H37 and H91 is reproducibly observed (Fig. 4b; also Supplementary Figs S3a, S3c and S4a). Together with the high efficiency of protein inactivation by H_2O_2 (Figs 3c and d), these results suggest that oxidation of either H37 or H91 is sufficient for derepression.

In summary, our findings reveal a previously unknown H_2O_2 -sensing mechanism by which PerR uses MCO reactions to regulate the expression of oxidative defence genes. Derepression of the PerR regulon is correlated with protein oxidation of histidine residues that are required as ligands for the regulatory metal ion (usually Fe^{2+}). No mechanisms are known for the repair of 2-oxo-histidine, and oxidized PerR may be degraded in the cell. The use of a sacrificial protein to regulate important defence responses is not unprecedented: both the LexA and Ada regulators are irreversibly modified in response to DNA damage^{18,19}. MCO reactions are known to

contribute to the toxicity of reactive oxygen species, to inactivate some enzymes and are implicated in both ageing^{20,21} and neurodegenerative diseases²². For example, 2-oxo-histidine formation upon exposure to H_2O_2 has been observed in human Cu,Zn superoxide dismutase¹⁴, an enzyme linked to familial amyotrophic lateral sclerosis, and in β -amyloid²³, a peptide implicated in Alzheimer's disease.

To our knowledge, MCO reactions in general, and histidine oxidation in particular, have not previously been observed to have a regulatory role in peroxide sensing. All characterized peroxide-sensing transcription factors, from both prokaryotes and eukaryotes, use reactive cysteine residues and, in most cases, oxidation leads to the reversible formation of disulphide bonds^{2,3,8}. Because the major genotoxic consequences of peroxide stress result from MCO reactions involving intracellular, loosely-chelated Fe, it is appropriate that the cell has adapted the same chemistry to sense oxidative stress¹. Unlike comparable thiol-based regulatory switches, such as OxyR, PerR is relatively insensitive to disulphide stress and the induction of peroxide-defence functions is tuned to both the availability of Fe^{2+} and increased intracellular peroxide levels.

METHODS

Purification of PerR:Zn and determination of metal content. PerR was purified after overexpression in *E. coli* BL21/DE3(pLysS) as previously described⁵ except that EDTA was added and thiol-reducing agents were omitted except where indicated. Briefly, the purification involved heparin-sepharose, Mono-Q, and Superdex-200 chromatography (Pharmacia). The purified protein retained high activity when stored at $-20^\circ C$ or $-80^\circ C$ in 20 mM TrisHCl pH 8.0, 100 mM NaCl, and 5% glycerol for up to two years. In general, 10 mM EDTA was critical for preventing metal-catalysed protein oxidation during purification. The concentration of PerR was determined using $\epsilon_{277\text{ nm}} = 10,400\text{ M}^{-1}\text{ cm}^{-1}$. PerR was purified to $>95\%$ homogeneity as judged by ESI-MS and SDS-polyacrylamide gel electrophoresis (PAGE) and contained ~ 0.8 Zn per monomer as judged by inductively coupled plasma mass spectrometry (ICP-MS).

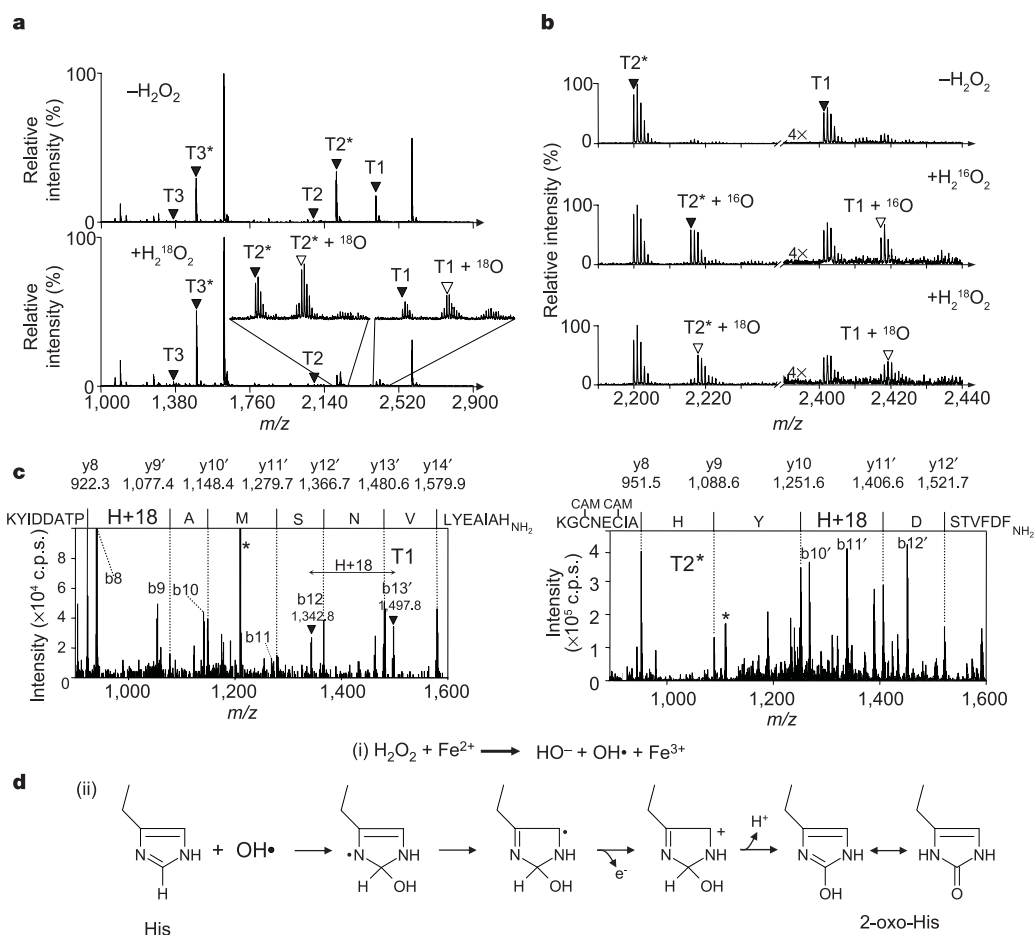


Figure 4 | Oxidation of PerR is localized to H37 and H91. **a**, Incorporation of ^{18}O (open arrows) into peptides T1 and T2* (Fig. 1a). **b**, Oxidation of PerR-FLAG by H_2O_2 *in vivo* leads to oxygen incorporation into T1 and T2*, but no cysteine oxidation (see extended spectra in Supplementary Fig. S2). Peak intensity for the right portion of the spectra is enlarged 4-fold (4 $\times$). Asterisks represent peptides containing carboxyamidomethylated cysteine residues. **c**, Localization of ^{18}O to H37 and H91 by ESI-MS. The $[T1 + 18]^{2+}$ and $[T2^* + 18]^{2+}$ parent ions (see asterisks on spectra) and carboxy-terminal (y-series) and amino-terminal (b-series) fragmentation products

are shown. For T1, y8 and y9' (and b12 and b13') differ by 155.1 Da (the molecular mass of histidine + ^{18}O), indicating oxidation of H37 (the prime symbol indicates +18 Da). For T2*, y10 and y11' differ by 155 Da, localizing ^{18}O to H91 (and not H93). (See Supplementary Figs S6 and S7 for additional details). m/z refers to the mass to charge ratio, and c.p.s., counts per second. **d**, Formation of 2-oxo-histidine by MCO¹⁶. The hydroxyl radical generated by the Fenton reaction (i) reacts rapidly with histidine (ii). Loss of an electron from the initial adduct may regenerate Fe^{2+} .

analysis. Under optimal purification conditions, PerR bound operator DNA with a dissociation constant (K_d) of ~ 1 nM as measured by either EMSA or FA and contained $\sim 60\%$ active molecules (Fig. 3c and d). For Fig. 3d, 50 μM PerR:Zn was reconstituted anaerobically with 50 μM Fe^{2+} and the amount of residual Fe^{2+} was determined using ferrozine²⁴ (Fz ; $\epsilon_{562\text{ nm}} = 27,900\text{ M}^{-1}\text{ cm}^{-1}$ for $Fe^{2+} - Fz_3$).

Electrophoretic mobility shift assays. PerR protein and a 250-base-pair ^{32}P -labelled *mrgA* promoter fragment (0.1–1.0 nM) were incubated in binding buffer (20 mM TrisHCl, pH 8.0, 5% glycerol, 5 $\mu g\text{ ml}^{-1}$ salmon sperm DNA, 50 $\mu g\text{ ml}^{-1}$ bovine serum albumin (BSA), 50 mM KCl, 100 μM $MnCl_2$) and separated by 6% PAGE with a 45 mM Tris-borate buffer containing 100 μM $MnCl_2$. For aerobic inactivation by Fe^{2+} (Fig. 2b), 1 μM PerR:Zn was incubated with 10 μM Fe^{2+} for 5 min at room temperature (22 $^\circ C$) with the appropriate concentration of Zn^{2+} or Mn^{2+} added for 5 min before or after Fe^{2+} as indicated. Diluted (20 $\times$) aliquots from these reactions were analysed by EMSA (final concentration of PerR:Zn was 50 nM). For Fig. 3c, PerR:Zn,Fe or PerR:Zn,Mn was added in twofold-increments from 0.23 to 30 nM and the K_d was estimated as the concentration of protein that led to $\sim 50\%$ binding.

Fluorescence anisotropy assays. A 6-carboxyfluorescein- (6F)-labelled *mrgA*-PerR box DNA fragment was generated by self-annealing of 5'-6F-CTAAATTA TAATAATTATAATTAG-3' (Integrated DNA Technologies). FA measurements ($\lambda_{ex} = 492\text{ nm}$, slit width = 15 nm; $\lambda_{em} = 520\text{ nm}$, slit width = 20 nm) were performed anaerobically in 3 ml of 20 mM TrisHCl pH 7.0, 5% glycerol, 100 mM KCl, with 100 nM DNA, 160 nM dimeric PerR:Zn (~ 100 nM active molecules), metal ions, H_2O_2 or EDTA, as indicated (Fig. 3b). Since near-quantitative inactivation of PerR:Zn was observed with 1 μM Fe and 1 μM H_2O_2 in 10 s,

we estimate $k_{inact} \approx 10^5\text{ M}^{-1}\text{ s}^{-1}$. For the determination of percentage activity, 10 nM DNA was titrated with PerR (up to 60 nM of dimer) in the presence of 100 μM Mn^{2+} (Fig. 3d). The anisotropy change versus PerR concentration was fitted using a 1:1 binding model.

Mass spectrometry. ESI-MS was used to measure protein oxidation (Fig. 2a) using a Bruker Esquire LC ion trap mass spectrometer. For standard acidic denaturing conditions, 25%(v/v) methanol and 1%(v/v) acetic acid in 5 mM ammonium bicarbonate was used. For the detection of bound Zn^{2+} (Fig. 2a), 5%(v/v) methanol in 5 mM ammonium bicarbonate was used. To monitor cysteine oxidation, samples were modified with 50 mM iodoacetamide in the presence of 50 mM EDTA to yield the carboxyamidomethyl derivative. Peptides containing cysteine-carboxyamidomethyl residues are indicated by asterisks in Fig. 4a and b.

Sites of oxidation were mapped by trypsin digestion followed by matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry using an Applied Biosystems 4700 mass spectrometer and further localized using ESI-MS on an Applied Biosystems 4000 Q Trap (Fig. 4 and Supplementary Figs S6 and S7). To monitor the incorporation of ^{18}O into the H37 and H91 peptides, 100 μM isotopically enriched $H_2^{18}O_2$ (90%) was used (Icon Isotopes).

Monitoring PerR oxidation *in vivo*. PerR activity was monitored using an *mrgA-cat-lacZ* fusion⁷ in cells grown in minimal medium with 100 μM iron and manganese as described previously⁷. Mutant alleles of *perR* (Fig. 1b and Supplementary Fig. S1) were generated using the QuikChange method (Stratagene).

To monitor PerR oxidation *in vivo*, we constructed a *perR*-FLAG fusion gene using pMUTIN-FLAG²⁵. The resulting gene was subcloned into pDG1730 (ref. 26) and integrated by double-crossover transformation into HB2080 (CU1065 *perR::kan amyE::cat*) to generate HB9620. The *perR*-FLAG construct was judged by complementation to be functional. Cells were grown in Luria-Bertani medium with or without 100 μ M H₂O₂ for 1 min, lysed with lysozyme and 2% TritonX-100 in 50 mM Tris, pH 7.4, 150 mM NaCl, 20 mM EDTA containing protease inhibitor (Roche Diagnostics). PerR-FLAG protein was recovered on an anti-FLAG M2 affinity gel (Sigma) followed by SDS-PAGE. Fractionated protein was analysed by in-gel tryptic digestion and mass spectrometry.

Received 16 November 2005; accepted 3 January 2006.

1. Imlay, J. A. Pathways of oxidative damage. *Annu. Rev. Microbiol.* **57**, 395–418 (2003).
2. Kiley, P. J. & Storz, G. Exploiting thiol modifications. *PLoS Biol.* **2**, e400 (2004).
3. Toledano, M. B., Delaunay, A., Monceau, L. & Tacnet, F. Microbial H₂O₂ sensors as archetypical redox signaling modules. *Trends Biochem. Sci.* **29**, 351–357 (2004).
4. Mongkolsuk, S. & Helmann, J. D. Regulation of inducible peroxide stress responses. *Mol. Microbiol.* **45**, 9–15 (2002).
5. Herbig, A. F. & Helmann, J. D. Roles of metal ions and hydrogen peroxide in modulating the interaction of the *Bacillus subtilis* PerR peroxide regulon repressor with operator DNA. *Mol. Microbiol.* **41**, 849–859 (2001).
6. Helmann, J. D. *et al.* The global transcriptional response of *Bacillus subtilis* to peroxide stress is coordinated by three transcription factors. *J. Bacteriol.* **185**, 243–253 (2003).
7. Fuangthong, M., Herbig, A. F., Bsat, N. & Helmann, J. D. Regulation of the *Bacillus subtilis* *fur* and *perR* genes by PerR: not all members of the PerR regulon are peroxide inducible. *J. Bacteriol.* **184**, 3276–3286 (2002).
8. Barford, D. The role of cysteine residues as redox-sensitive regulatory switches. *Curr. Opin. Struct. Biol.* **14**, 679–686 (2004).
9. Hahn, J. S., Oh, S. Y., Chater, K. F., Cho, Y. H. & Roe, J. H. H₂O₂-sensitive Fur-like repressor CatR regulating the major catalase gene in *Streptomyces coelicolor*. *J. Biol. Chem.* **275**, 38254–38260 (2000).
10. Ortiz de Orue Lucana, D., Troller, M. & Schrempf, H. Amino acid residues involved in reversible thiol formation and zinc ion binding in the *Streptomyces reticuli* redox regulator FurS. *Mol. Genet. Genomics* **268**, 618–627 (2003).
11. Pohl, E. *et al.* Architecture of a protein central to iron homeostasis: crystal structure and spectroscopic analysis of the ferric uptake regulator. *Mol. Microbiol.* **47**, 903–915 (2003).
12. Chen, L., Keramati, L. & Helmann, J. D. Coordinate regulation of *Bacillus subtilis* peroxide stress genes by hydrogen peroxide and metal ions. *Proc. Natl Acad. Sci. USA* **92**, 8190–8194 (1995).
13. Lee, C. *et al.* Redox regulation of OxyR requires specific disulfide bond formation involving a rapid kinetic reaction path. *Nat. Struct. Mol. Biol.* **11**, 1179–1185 (2004).
14. Uchida, K. & Kawakishi, S. Identification of oxidized histidine generated at the active site of Cu,Zn-superoxide dismutase exposed to H₂O₂: Selective generation of 2-oxo-histidine at the histidine 118. *J. Biol. Chem.* **269**, 2405–2410 (1994).
15. Uchida, K. & Kawakishi, S. 2-Oxo-histidine as a novel biological marker for oxidatively modified proteins. *FEBS Lett.* **332**, 208–210 (1993).
16. Schoneich, C. Mechanisms of metal-catalyzed oxidation of histidine to 2-oxo-histidine in peptides and proteins. *J. Pharm. Biomed. Anal.* **21**, 1093–1097 (2000).
17. Jewett, S. L., Rocklin, A. M., Ghanevati, M., Abel, J. M. & Marach, J. A. A new look at a time-worn system: oxidation of CuZn-SOD by H₂O₂. *Free Radic. Biol. Med.* **26**, 905–918 (1999).
18. He, C. *et al.* A methylation-dependent electrostatic switch controls DNA repair and transcriptional activation by *E. coli* Ada. *Mol. Cell* **20**, 117–129 (2005).
19. Luo, Y. *et al.* Crystal structure of LexA: a conformational switch for regulation of self-cleavage. *Cell* **106**, 585–594 (2001).
20. Balaban, R. S., Nemoto, S. & Finkel, T. Mitochondria, oxidants, and aging. *Cell* **120**, 483–495 (2005).
21. Nystrom, T. Role of oxidative carbonylation in protein quality control and senescence. *EMBO J.* **24**, 1311–1317 (2005).
22. Wright, A. F. *et al.* Lifespan and mitochondrial control of neurodegeneration. *Nature Genet.* **36**, 1153–1158 (2004).
23. Schoneich, C. & Williams, T. D. Cu(II)-catalyzed oxidation of beta-amyloid peptide targets His13 and His14 over His6: Detection of 2-Oxo-histidine by HPLC-MS/MS. *Chem. Res. Toxicol.* **15**, 717–722 (2002).
24. Berlett, B. S., Levine, R. L., Chock, P. B., Chevion, M. & Stadtman, E. R. Antioxidant activity of Ferrozine-iron-amino acid complexes. *Proc. Natl Acad. Sci. USA* **98**, 451–456 (2001).
25. Kaltwasser, M., Wiegert, T. & Schumann, W. Construction and application of epitope- and green fluorescent protein-tagging integration vectors for *Bacillus subtilis*. *Appl. Environ. Microbiol.* **68**, 2624–2628 (2002).
26. Guérout-Fleury, A.-M., Frandsen, N. & Stragier, P. Plasmids for ectopic integration in *Bacillus subtilis*. *Gene* **180**, 57–61 (1996).
27. Schwede, T., Kopp, J., Güex, N. & Peitsch, M. C. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **31**, 3381–3385 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Fuangthong and S. Soonsanga for their initial analyses of functionally altered *perR* variants, E. Madsen for help with the anaerobic experiments, S. Zhang and R. Sherwood for assistance with mass spectrometry, and G. Storz, P. Kiley, J. Imlay, and C. M. Moore for comments. This work was supported by grants from the NSF and NIH.

Author Contributions: J.-W.L. performed all experimental work, J.-W.L. and J.D.H. analysed results and co-wrote the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.D.H. (jdh9@cornell.edu).

LETTERS

Crystal structure of the non-haem iron halogenase SyrB2 in syringomycin biosynthesis

Leah C. Blasiak¹, Frédéric H. Vaillancourt^{2†}, Christopher T. Walsh² & Catherine L. Drennan¹

Non-haem Fe(II)/ α -ketoglutarate (α KG)-dependent enzymes harness the reducing power of α KG to catalyse oxidative reactions, usually the hydroxylation of unactivated carbons, and are involved in processes such as natural product biosynthesis, the mammalian hypoxic response, and DNA repair^{1,2}. These enzymes couple the decarboxylation of α KG with the formation of a high-energy ferryl-oxo intermediate that acts as a hydrogen-abstracting species^{2–4}. All previously structurally characterized mononuclear iron enzymes contain a 2-His, 1-carboxylate motif that coordinates the iron^{1,2}. The two histidines and one carboxylate, known as the ‘facial triad’, form one triangular side of an octahedral iron coordination geometry. A subclass of mononuclear iron enzymes has been shown to catalyse halogenation reactions, rather than the more typical hydroxylation reaction^{5,6}. SyrB2, a member of this subclass, is a non-haem Fe(II)/ α KG-dependent halogenase that catalyses the chlorination of threonine in syringomycin E biosynthesis⁵. Here we report the structure of SyrB2 with both a chloride ion and α KG coordinated to the iron ion at 1.6 Å resolution. This structure reveals a previously unknown coordination of iron, in which the carboxylate ligand of the facial triad is replaced by a chloride ion.

The gene encoding SyrB2 is found in the non-ribosomal peptide synthetase gene cluster for production of the phytotoxin syringomycin E in *Pseudomonas syringae* pv. *syringae* B301D⁷ (Fig. 1a). SyrB2 catalyses the Fe(II)-, α KG-, Cl[−]- and O₂-dependent conversion of L-Thr to 4-Cl-L-Thr (ref. 5 and Fig. 1b), a modification essential for the antifungal activity of the natural product⁸. SyrB2 is active on L-Thr only when the amino acid is tethered to the phosphopantetheine prosthetic arm of the holo-SyrB1 thiolation domain⁵. Another class of halogenases, the flavin-dependent enzymes, have been structurally characterized and shown to catalyse halogenation at aromatic or electron-rich carbons⁹. Halogenation at aliphatic carbons, as in the natural products syringomycin E (ref. 7), a phytotoxin, barbamide¹⁰, a molluscicidal compound, and victorin¹¹, a host-selective crop toxin, requires a more reactive catalyst. It has been shown that Fe(II)/ α KG-dependent enzymes such as SyrB2 catalyse these more challenging halogenation reactions^{5,6}. Other Fe(II)/ α KG-dependent halogenases include BarB1/BarB2 and CmaB, which are involved in the biosynthesis of barbamide¹⁰ and the phytotoxin coronatine⁶, respectively.

The core of SyrB2 is composed of an eight-stranded anti-parallel β -sandwich ‘jelly-roll’ motif (comprising strands β 2, β 12, β 5, β 10 and strands β 9, β 6, β 11, β 3) that defines the cupin superfamily (Fig. 2a and Supplementary Fig. 1). A structural alignment search for SyrB2 using the DALI server¹² produced five matches with Z-scores >4, all members of the cupin superfamily. The closest structural relative, with a Z-score of 8.5, is human factor-inhibiting hypoxia-inducible factor-1 (FIH-1), an Fe(II)/ α KG-dependent asparaginyl

hydroxylase involved in the mammalian hypoxic response¹³. SyrB2 is the first structurally characterized mononuclear iron protein that does not contain the classic 2-His, 1-carboxylate iron coordination motif. The two histidine ligands (His 116 and His 235) are present, but the iron has no carboxylate ligand (Fig. 2b). Instead, this site is occupied by a chloride ion in our structure, as judged by an 8 σ peak in the $F_o - F_c$ map.

To verify that this iron ligand is a halide ion, we took advantage of the ability of SyrB2 to catalyse bromination¹⁴ as well as chlorination⁵. Unlike chloride, which does not have a strong anomalous signal, bromide has an anomalous peak at 0.9195 Å that is well separated from the iron anomalous peak at 1.7397 Å. Bromide-soaked crystals were prepared and a dispersive difference Fourier map was calculated between the iron and bromide edges, confirming the location of the

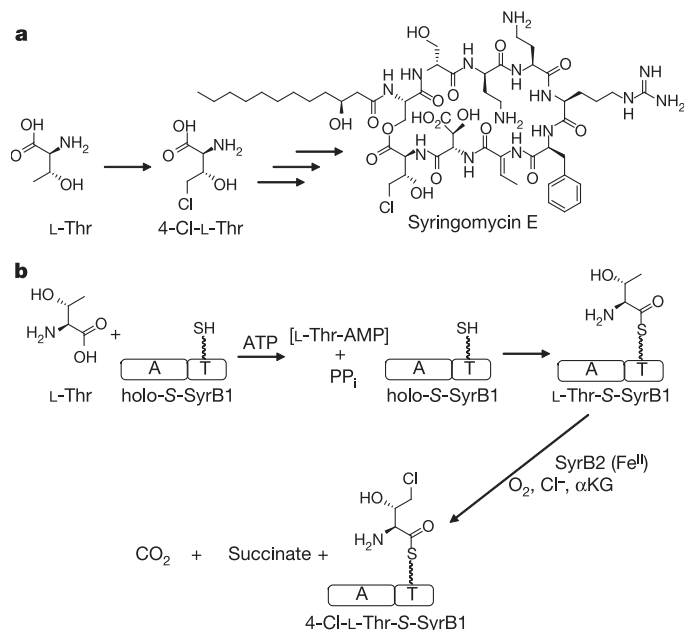


Figure 1 | SyrB2 activity in syringomycin E biosynthesis. **a**, SyrB2 catalyses the conversion of L-Thr to 4-Cl-L-Thr, which is then incorporated into syringomycin E. **b**, Scheme showing the adenylation of L-Thr by SyrB1, loading of L-Thr-AMP onto the SyrB1 phosphopantetheine group, and subsequent chlorination by SyrB2. The phosphopantetheine prosthetic group on SyrB1, indicated by a wavy line, is posttranslationally attached through a serine residue and functions as a ‘swinging arm’ for transfer of the covalently attached amino acid between enzyme active sites. A, adenylation domain; T, thiolation domain.

¹Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ²Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA. [†]Present address: Department of Biological Sciences, Research and Development, Boehringer Ingelheim (Canada) Ltd, Laval, Quebec H7S 2G5, Canada.

Table 1 | Activity and iron content of SyrB2 variants

	Wild-type SyrB2	A118D SyrB2	A118E SyrB2
Activity	+	—	—
Iron content	83.0%	65.3%	74.4%

iron and halide ion (Fig. 2d). Another bromide ion is located on the surface of each SyrB2 molecule, but it is unlikely to have biological significance.

In addition to the two histidines and one chloride ligand, the remaining iron coordination sites are occupied by the α KG and a water molecule. The water molecule binds opposite His 235 at the site where dioxygen would probably bind during catalysis (Fig. 2b). α KG coordinates iron in a bidentate fashion, with the C1 carboxylate opposite His 116, and the C2 ketone opposite the chloride ligand. The α KG is anchored in the active site through hydrogen bonds between the C5 carboxylate and Thr 113, Arg 248 and Trp 145 (Fig. 2c). Bidentate binding of α KG and the arginine and threonine hydrogen-bonding interactions are conserved across many members of the Fe(II)/ α KG-dependent family², but the interaction of Trp 145 is unique to SyrB2. Phe 104 stacks above the planar α KG, whereas Arg 254 is situated 2.9 Å above the planar α KG and may be important in positioning the C1 carboxylate. A similarly placed arginine or asparagine is observed in many other family members².

The chloride ligand in SyrB2 sits in a largely hydrophobic pocket composed of residues Ala 118, Phe 121 and the β -carbon of Ser 231.

However, it also forms interactions with two water molecules, which in turn form hydrogen bonds with Thr 143 and Asn 123 (Fig. 2c). The interactions between the chloride ligand and water may be involved in lengthening the Fe–Cl bond, which at 2.44 Å is slightly longer than the average six-coordinate Fe(II)–Cl bond length of 2.30 Å found in the Cambridge Structural Database¹⁵ and other model compounds¹⁶. All other iron–ligand bond lengths are typical of Fe(II)/ α KG-dependent enzymes.

The position of Ala 118 in SyrB2 seems to have a role in the protein's halogenase activity by creating space in the active site for the chloride ligand to bind iron. Ala 118 is completely conserved among known mononuclear iron halogenases (Supplementary Fig. 2a). A structure-based sequence alignment of SyrB2 with non-halogenase Fe(II)/ α KG-dependent enzymes shows that the active site residues are located at structurally equivalent positions (Supplementary Fig. 2b). Ala 118 in SyrB2 replaces the conserved aspartate or glutamate that serves as an iron ligand in other mononuclear iron enzymes. An alignment of SyrB2 with its most structurally similar relative, FIH-1, shows replacement of the aspartate by alanine, creating space for chloride to bind (Fig. 3a). If a glutamate is modelled in place of Ala 118 in the SyrB2 structure, it is ideally positioned for iron coordination. To verify that substitution of Ala 118 with aspartate or glutamate would abolish chlorination activity while retaining iron-binding ability, A118D and A118E variants were expressed, purified and reconstituted with Fe(II) and α KG. Both variants bound Fe(II), but the mutations completely abrogated chlorination activity (Table 1 and Supplementary Fig. 3).

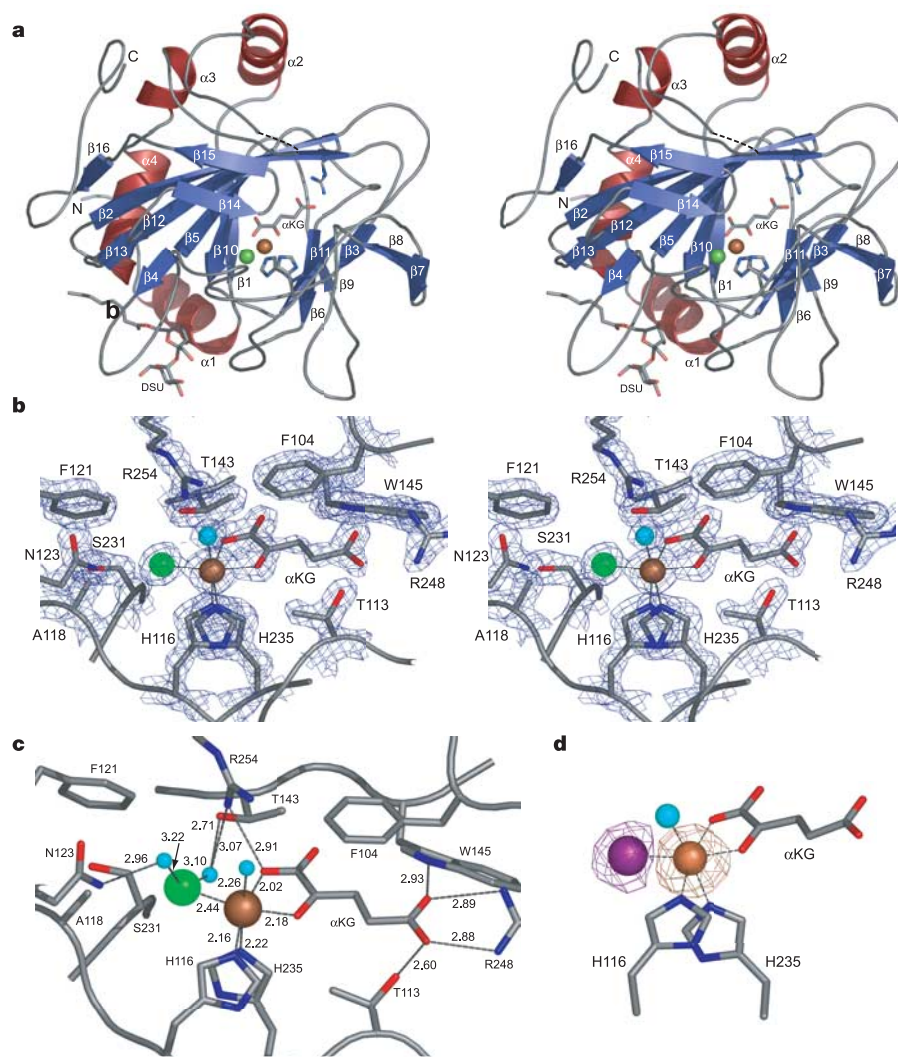


Figure 2 | Structure of SyrB2 with bound iron, chloride and α KG. Iron is shown in brown, chloride in green. **a**, Stereo view of the SyrB2 structure. Two protein ligands of iron (His 116 and His 235), α KG and Arg 248 are shown in stick representation. A disordered loop (residues 56–60) is indicated by the broken line. A detergent molecule (DSU) from the crystallization condition packs along the bottom surface of the structure. **b**, Stereo view of the SyrB2 active site. A $2F_o - F_c$ composite omit map contoured to 2σ is shown in blue mesh. **c**, SyrB2 active site distances. **d**, Dispersive difference Fourier map from bromide-containing crystals, calculated by subtracting data collected at 0.9197 Å (bromide edge) from data collected at 1.7340 Å (iron edge). The purple mesh ($+4\sigma$) confirms the location of the halide ion; the brown mesh (-4σ) confirms the location of the iron.

Thus, in SyrB2, it seems that nature chooses alanine at this position instead of an aspartate or glutamate residue to create an iron coordination site for chloride binding.

Despite the requirement for a 'facial triad' of ligands in all other Fe(II)/ α KG dependent hydroxylases, it seems unlikely that a carboxylate ligand could bind to iron in SyrB2 in any conformation or stage of catalysis. The closest candidates (Asp 114 and Asp 119) are >10 Å from the iron, pointing in the opposite direction (Fig. 3a). Asp 114 forms part of the β -sandwich core, and Asp 119 is not conserved. The iron coordination structure with only two direct protein ligands also helps to explain the low affinity of apo-SyrB2 for Fe(II) during reconstitution, which yields only $\sim 35\%$ incorporation, as opposed to reconstitution with α KG followed by Fe(II), which yields $\sim 90\%$ incorporation⁵.

So far, SyrB2 crystallizes only in the presence of α KG, indicating that this cofactor provides structural stability to the protein. The structure of SyrB2 with only α KG bound has two channels leading to the active site (Fig. 3b). The loops covering the active site in this structure have higher *B*-factors (by 10 – 20 Å²) than the average for the rest of the structure, indicating that they have some degree of mobility. Once iron and chloride bind, the protein closes down around the active site (Fig. 3c). This closure involves motions of two loops and a β -hairpin (Fig. 3d). The substrate of SyrB2, L-Thr, which is loaded on the phosphopantetheine arm of the 66-kDa SyrB1 protein, probably enters the active site through the remaining narrow tunnel above α KG (Fig. 3c). The distance from the iron to the edge of the tunnel is ~ 17 Å, similar to the length of the extended phosphopantetheine arm (~ 20 Å). Access through this tunnel is limited by the side chain of Phe 196 (Fig. 3c and Supplementary Fig. 4), which may be involved in gating the entrance to the active site. We cannot rule out the possibility, however, that SyrB2 undergoes a conformational

change on interaction with SyrB1 that opens a different passageway to the active site.

In predictions of how the Fe(II)/ α KG-dependent enzyme scaffold might be used for halogenation, a persistent issue was the location of the chloride. Without a structure, it was presumed that SyrB2 would contain the characteristic HX(D/E) X_n H motif for iron binding. The sequence of SyrB2 does contain a conserved HXD motif, but it is located >15 Å from the active site in our structure. If we assumed three protein ligands and bidentate α KG, only one coordination site for dioxygen would remain, leaving none for chloride binding. If this were true, chloride binding to iron would have to take place after decarboxylation of α KG (and subsequent release of succinate or CO₂) and formation of the ferryl-oxo intermediate. However, the high-energy ferryl-oxo species is expected to be very reactive; thus, it would be unfavourable to form this intermediate before chloride were in position to react. The missing carboxylate ligand in SyrB2 elegantly solves this mechanistic dilemma. Our crystallographic data confirm that halide ions bind directly to iron, before decarboxylation of α KG and formation of a reactive ferryl-oxo species.

Based on the structure of the resting state of SyrB2 and by analogy with the Fe(II)/ α KG-dependent hydroxylation mechanism of TauD¹⁷, we propose a working mechanism for halogenation by SyrB2 (Fig. 4). As observed crystallographically, α KG, chloride and water coordinate the iron in the resting Fe(II) state (Fig. 4, species A). L-Thr-S-SyrB1 binding would then exclude water from the active site and allow dioxygen to bind (Fig. 4, species B and C). Decarboxylation of α KG would lead to the formation of a high-energy ferryl-oxo intermediate (Fig. 4, species D), which would then abstract a hydrogen atom from the substrate. The substrate radical would abstract the chlorine atom (Fig. 4, species E), producing chlorinated L-Thr-S-SyrB1 and regenerating the reduced Fe(II) centre. After formation of the substrate carbon radical, however, some competition between transfer of Cl $\cdot$ and OH $\cdot$ would be expected. No threonine hydroxylation side reaction has been detected, indicating that Cl $\cdot$ transfer is greatly favoured⁵. There are several possible explanations for this selectivity, including the positioning of the substrate in the active site during the reaction, and the lower potential of Cl $\cdot$ versus OH $\cdot$, as has been proposed to explain similar reactivity in model complexes^{5, 18}.

The crystal structure of SyrB2 reveals an unexpected iron scaffold that was not predicted from the protein sequence and provides insight into the function of Fe(II)/ α KG-dependent halogenases. SyrB2 could serve as an excellent model system for studying the

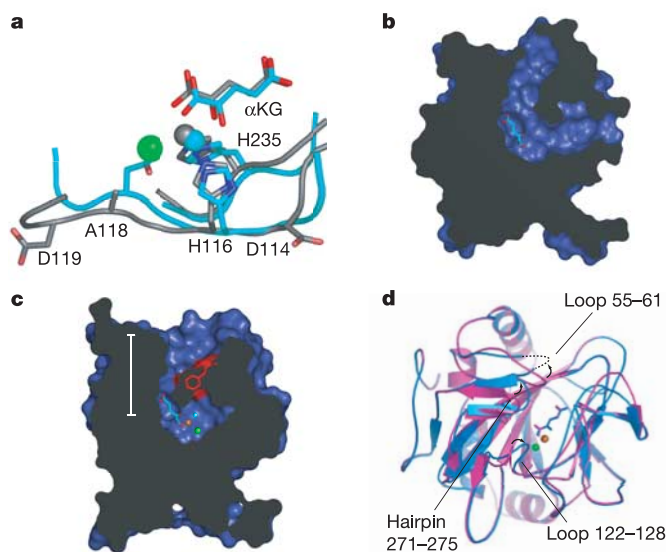


Figure 3 | Structural comparisons. **a**, View of the active site from a structural alignment of SyrB2 (grey) and FIH-1 (cyan). Alignment was done with LSQMAN²⁹. Ninety-nine residues (33% of SyrB2) align with a root mean square deviation of 1.95 Å. Iron and chloride from SyrB2 are shown in grey and green, respectively; iron from FIH-1 is shown in cyan. Labelled amino acids correspond to SyrB2. **b**, Cut-away view of SyrB2 with bound α KG and no iron. Two channels lead to the active site. Phe 196 is not visible owing to clipping planes. **c**, Cut-away view of Cl-Fe(II)- α KG-SyrB2 in a similar orientation to **b**, but the molecular surface is clipped to show Phe 196 (red) clearly. SyrB2 closes down around the bound iron (brown) and chloride (green), leaving only one opening to the active site. The water molecule located at the predicted dioxygen-binding site is shown in cyan. Scale bar, ~ 15 Å. **d**, α KG-SyrB2 (magenta) and Cl-Fe(II)- α KG-SyrB2 (blue). Arrows show change in conformation on binding of iron and chloride.

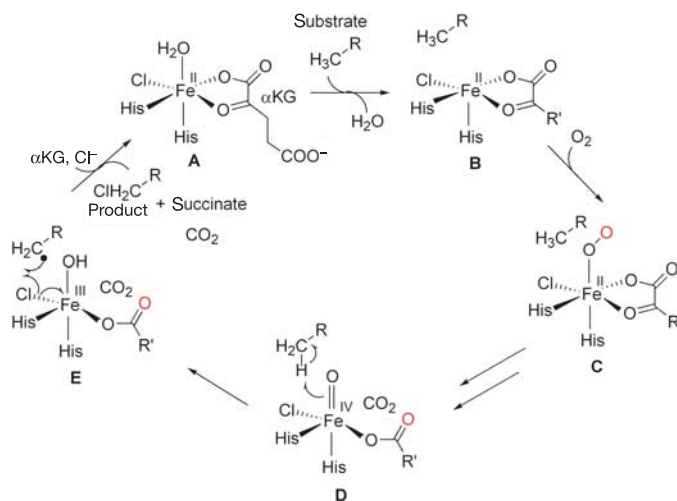


Figure 4 | Proposed mechanism for the reaction catalysed by SyrB2. See text for a description of the proposed steps. The crystallized form of SyrB2 is labelled A. The oxygen atom inserted into α KG is shown in red.

reactivity of iron halides and the detailed mechanism of halogenation by mononuclear iron enzymes.

METHODS

Protein purification, crystallization, data collection, site-directed mutagenesis and assays of the A118D and A118E variants were done as described in Supplementary Methods.

Structure determination. Phases were determined from a SAD experiment on seleno-methionine-labelled SyrB2 with α KG. Sixteen selenium sites (eight per molecule) were found in SOLVE¹⁹ and refined in SHARP²⁰, with an acentric figure of merit of 0.431 to 2.3 Å resolution. To improve the quality of the electron density map, non-crystallographic averaging (over the two molecules in the asymmetric unit) and solvent flattening were done in DM²¹. The initial model was built manually using XFIT²² and refined in CNS²³ with non-crystallographic symmetry (NCS) restraints applied. Several rounds of iterative model building and refinement without NCS restraints followed. When the *R*-factors dropped below 25.5% (work) and 28.0% (free), the model was refined against the high-resolution α KG–SyrB2 and Fe(II) data sets. The same set of test reflections (extended for higher resolution structures) was used to calculate *R*_{free} for all models. Final rounds of refinement for all structures were done with the program REFMAC5 (ref. 24). Simulated annealing composite omit maps calculated in CNS were used to verify structures throughout refinement.

The final model of α KG–SyrB2 contains residues 2–310 in molecule A, and residues 3–55, 61–133 and 139–310 in molecule B out of 310 residues. The final models of Cl/Br–Fe(II)– α KG–SyrB2 contain residues 3–56 and 61–310 in molecule A, and residues 4–53 and 62–310 in molecule B. All structures contain one detergent molecule from the crystallization condition (*n*-decanoyl sucrose; DSU), bound on the surface of molecule A. The topology and parameter files for α KG were obtained from the REFMAC5 (ref. 24) dictionary, and the DSU dictionary was generated in the program SKETCHER²⁵. Planar restraints on the sugar rings were removed for refinement. A summary of the final refinement statistics is available in Supplementary Table S1. Assessment of the models in PROCHECK²⁶ found no unfavourable geometries and >90% of residues were in the ‘most favoured’ conformation. A dispersive difference Fourier map was calculated in CNS by subtracting the bromide inflection data from the iron inflection data using phases calculated from a model that contained only protein atoms. The observed negative dispersive difference electron density confirmed the presence of one iron site per asymmetric unit, and the positive dispersive difference electron density indicated the presence of two bromide sites (one bound to iron in the active site and one on the surface) per molecule. Secondary structure assignment was done in DSSP²⁷, and figures were generated in PyMOL²⁸.

Received 3 November; accepted 22 December 2005.

- Koehnert, K. D., Emerson, J. P. & Que, L. Jr. The 2-His-1-carboxylate facial triad: a versatile platform for dioxygen activation by mononuclear non-heme iron(II) enzymes. *J. Biol. Inorg. Chem.* **10**, 87–93 (2005).
- Hausinger, R. P. Fe(II)/ α -ketoglutarate-dependent hydroxylases and related enzymes. *Crit. Rev. Biochem. Mol. Biol.* **39**, 21–68 (2004).
- Price, J. C., Barr, E. W., Hoffart, L. M., Krebs, C. & Bollinger, J. M. Jr. Kinetic dissection of the catalytic mechanism of taurine: α -ketoglutarate dioxygenase (TauD) from *Escherichia coli*. *Biochemistry* **44**, 8138–8147 (2005).
- Price, J. C., Barr, E. W., Tirupati, B., Bollinger, J. M. Jr & Krebs, C. The first direct characterization of a high-valent iron intermediate in the reaction of an α -ketoglutarate-dependent dioxygenase: a high-spin Fe^{IV} complex in taurine/ α -ketoglutarate dioxygenase (TauD) from *Escherichia coli*. *Biochemistry* **42**, 7497–7508 (2003).
- Vaillancourt, F. H., Yin, J. & Walsh, C. T. SyrB2 in syringomycin E biosynthesis is a non-heme Fe^{II} α -ketoglutarate and O₂ dependent halogenase. *Proc. Natl Acad. Sci. USA* **102**, 10111–10116 (2005).
- Vaillancourt, F. H., Yeh, E., Vosburg, D. A., O'Connor, S. E. & Walsh, C. T. Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosynthesis. *Nature* **436**, 1191–1194 (2005).
- Guenzi, E., Galli, G., Grgurina, I., Gross, D. C. & Grandi, G. Characterization of the syringomycin synthetase gene cluster. A link between prokaryotic and eukaryotic peptide synthetases. *J. Biol. Chem.* **273**, 32857–32863 (1998).
- Grgurina, I. et al. Relevance of chlorine-substituent for the antifungal activity of syringomycin and syringotoxin, metabolites of the phytopathogenic bacterium *Pseudomonas syringae* pv. *syringae*. *Experientia* **50**, 130–133 (1994).
- Dong, C. et al. Tryptophan 7-halogenase (PrnA) structure suggests a mechanism for regioselective chlorination. *Science* **309**, 2216–2219 (2005).
- Chang, Z. et al. The barbamide biosynthetic gene cluster: a novel marine cyanobacterial system of mixed polyketide synthase (PKS)–non-ribosomal peptide synthetase (NRPS) origin involving an unusual trichloroleucyl starter unit. *Gene* **296**, 235–247 (2002).
- Walton, J. D. Host-selective toxins: agents of compatibility. *Plant Cell* **8**, 1723–1733 (1996).
- Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595–603 (1996).
- Dann, C. E., Bruick, R. K. & Deisenhofer, J. Structure of factor-inhibiting hypoxia-inducible factor 1: an asparaginyl hydroxylase involved in the hypoxic response pathway. *Proc. Natl Acad. Sci. USA* **99**, 15351–15356 (2002).
- Vaillancourt, F. H., Vosburg, D. A. & Walsh, C. T. Dichlorination and bromination of a threonyl-S-carrier protein by the nonheme Fe^{II} halogenase SyrB2. *ChemBioChem* (in the press).
- Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. B* **58**, 380–388 (2002).
- Goldsmith, C. R., Jonas, R. T., Cole, A. P. & Stack, D. P. A spectrochemical walk: single-site perturbation within a series of six-coordinate ferrous complexes. *Inorg. Chem.* **41**, 4642–4652 (2002).
- Grzyska, P. K. et al. Steady-state and transient kinetic analyses of taurine/ α -ketoglutarate dioxygenase: effects of oxygen concentration, alternative sulfonates, and active-site variants on the Fe^{IV}-oxo intermediate. *Biochemistry* **44**, 3845–3855 (2005).
- Kojima, T., Leising, R. A., Yan, S. & Que, L. Jr. Alkane functionalization at nonheme iron center. Stoichiometric transfer of metal-bound ligands to alkane. *J. Am. Chem. Soc.* **115**, 11328–11335 (1993).
- Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
- La Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement in MIR and MAD methods. *Methods Enzymol.* **276**, 472–494 (1997).
- Cowan, K. DM: an automated procedure for phase improvement by density modification. *CCP4/ESF-EACBM Newslett. Protein Crystallogr.* **31**, 34–38 (1994).
- McRee, D. E. XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165 (1999).
- Brunger, A. T. et al. Crystallography and NMR system (CNS): a new software system for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr. D* **59**, 1131–1137 (2003).
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK—a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
- Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
- Delano, W. L. The PyMOL Molecular Graphics System (<http://www.pymol.org/>) (2002).
- Kleywegt, G. J. & Jones, T. A. A super position. *CCP4/ESF-EACBM Newslett. Protein Crystallogr.* **31**, 9–14 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. W. Nix and T. Doukov for help with data collection; and M. Fischbach for critically reading the manuscript. The Advanced Light Source and Stanford Synchrotron Radiation Laboratory are supported by the US Department of Energy. The SSRL Structural Molecular Biology Program is also supported by the National Institutes of Health. This work was supported in part by grants from the NIH (to C.L.D., to C.T.W. and to L.C.B.), a Merck-sponsored Fellowship of the Helen Hay Whitney Foundation (to F.H.V.) and a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship (to F.H.V.).

Author Information The coordinates and structure factors for α KG–SyrB2, Cl–Fe(II)– α KG–SyrB2 and Br–Fe(II)– α KG–SyrB2 have been deposited in the Protein Data Bank with accession codes 2FCT, 2FCU and 2FCV, respectively. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.L.D. (cdrennan@MIT.EDU).

LETTERS

Structural basis for the spectral difference in luciferase bioluminescence

Toru Nakatsu^{1,2}, Susumu Ichiyama¹, Jun Hiratake³, Adrian Saldanha³, Nobuyuki Kobashi¹, Kanzo Sakata³ & Hiroaki Kato^{1,2}

Fireflies communicate with each other by emitting yellow-green to yellow-orange brilliant light. The bioluminescence reaction, which uses luciferin, Mg-ATP and molecular oxygen to yield an electronically excited oxyluciferin species, is carried out by the enzyme luciferase. Visible light is emitted during relaxation of excited oxyluciferin to its ground state. The high quantum yield¹ of the luciferin/luciferase reaction and the change in bioluminescence colour caused by subtle structural differences in luciferase have attracted much research interest. In fact, a single amino acid substitution in luciferase changes the emission colour from yellow-green to red^{2–5}. Although the crystal structure of luciferase from the North American firefly (*Photinus pyralis*) has been described^{6,7}, the detailed mechanism for the bioluminescence colour change is still unclear^{8–11}. Here we report the crystal structures of wild-type and red mutant (S286N) luciferases from the Japanese Genji-botaru (*Luciola cruciata*) in complex with a high-energy intermediate analogue, 5'-O-[N-(dehydroluciferyl)-sulfamoyl]adenosine (DLSA). Comparing these structures to those of the wild-type luciferase complexed with AMP plus oxyluciferin (products) reveals a significant conformational change in the wild-type enzyme but not in the red mutant. This conformational change involves movement of the hydrophobic side chain of Ile 288 towards the benzothiazole ring of DLSA. Our results indicate that the degree of molecular rigidity of the excited state of oxyluciferin, which is controlled by a transient movement of Ile 288, determines the colour of bioluminescence during the emission reaction.

The bioluminescence reaction catalysed by firefly luciferase proceeds via the initial formation of an enzyme-bound luciferyl adenylate intermediate¹² (Fig. 1a). This mode of substrate activation is commonly used by AMP-forming ligases such as aminoacyl-tRNA synthetases¹³, acyl-CoA ligases¹⁴ and nonribosomal peptide synthetases¹⁵. Luciferases share significant primary sequence similarities with acyl-CoA ligases¹⁶ and nonribosomal peptide synthetases¹⁷, but not with aminoacyl-tRNA synthetases. We synthesized DLSA (Fig. 1b) as a ligand for X-ray structural analysis of firefly luciferase, because we assumed that the first half of the reaction catalysed by luciferase is chemically equivalent to that of the amino acylation catalysed by aminoacyl-tRNA synthetases. DLSA is analogous to 5'-O-[N-(aminoacyl)-sulfamoyl]adenosine, a potent inhibitor of aminoacyl-tRNA synthetases¹⁸, and was expected to serve as a stable analogue of the luciferyl adenylate intermediate¹⁹ (Supplementary Methods).

The crystal structure of the wild-type luciferase from *Luciola cruciata* (LcrLuc(WT)) complexed with DLSA was determined at 1.3 Å resolution (Fig. 2a) by the molecular replacement method using the unliganded crystal structure of luciferase from *Photinus*

pyralis (PpyLuc) (Protein Data Bank accession number 1LCI) as a model⁶. The overall structure of LcrLuc(WT) complexed with DLSA consists of a large amino-terminal domain and a small carboxy-terminal domain that are connected by a flexible linker loop (residues 438–442) (Fig. 2a). The spatial arrangement of the N- and C-terminal domains in the LcrLuc(WT)·DLSA complex structure is different from that in the PpyLuc structure⁶, but is similar to those in nonribosomal peptide synthetases structures (for example, PheA in complex with substrates²⁰ and DhBE with and without substrates¹⁷).

In the structure of the LcrLuc(WT)·DLSA complex, the DLSA molecule gives a clear electron density in the active site (Fig. 2b). The dehydroluciferin moiety adopts a *trans* form with a rotational angle of ~7° with respect to the C2–C2' bond, and is bound in a hydrophobic pocket consisting of α8 (amino acid residues 248–260), β12 (286–289), β13 (313–316), β14 (339–342), β15 (351–353) and a loop (343–350), while the entrance of the pocket is blocked by the adenosine moiety (Fig. 2a). The benzothiazole ring of DLSA is in van der Waals contact with the side chains of Phe 249, Thr 253, Ile 288 and Ala 350, and with the main chains of the β13 and β14 strands. A water molecule, Wat 1, is hydrogen-bonded to N3' of DLSA (2.83 Å) and Ser 349 Oγ (2.70 Å). Details of active-site residues are shown in Fig. 2c (see also Supplementary Fig. S1a).

We also solved the structures of LcrLuc(WT) in complex with Mg-ATP (reactant) and with AMP and oxyluciferin (products) at 2.3 and 1.6 Å resolution, respectively. The conformations of the dehydroluciferin and sulfamoyladenine moieties in the DLSA complex are very similar to those of oxyluciferin and AMP, respectively, in the AMP/oxyluciferin complex (Fig. 3a), indicating that the intermediate analogue DLSA is a fairly good mimic of oxyluciferin and AMP. Assuming that the structures of LcrLuc(WT) in complex with Mg-ATP (reactant), DLSA (high-energy intermediate) and AMP/oxyluciferin (products) represent snapshots along the enzymatic reaction coordinate, differences in the active-site structures around the ligand may reflect catalytically relevant movement of the amino acid residues.

The structure of the Mg-ATP complex is essentially the same as that of the AMP/oxyluciferin complex (see Supplementary Fig. S1b), but significant differences are observed in the structure of the DLSA complex (Fig. 3a). The β12 strand in the DLSA complex is much closer to the luciferin-binding site than in the other structures. In particular, the Cα of Ile 288 is 1.5 Å closer to the DLSA molecule, and the side chain of Ile 288 is also closer to DLSA and is rotated by 131°. This movement seems to be linked to the switching of a hydrogen-bonding network involving the side chain of Ser 286: Ser 286 is hydrogen-bonded to Glu 313 in LcrLuc(WT) complexed with AMP plus oxyluciferin, but is no longer hydrogen-bonded to Glu 313 in the

¹Kinetic Crystallography Research Team, Membrane Dynamics Research Group, RIKEN Harima Institute at SPring-8, 1-1-1 Kouto, Mikazuki-cho, Sayo-gun, Hyogo 679-5148, Japan. ²Department of Structural Biology, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan.

³Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan.

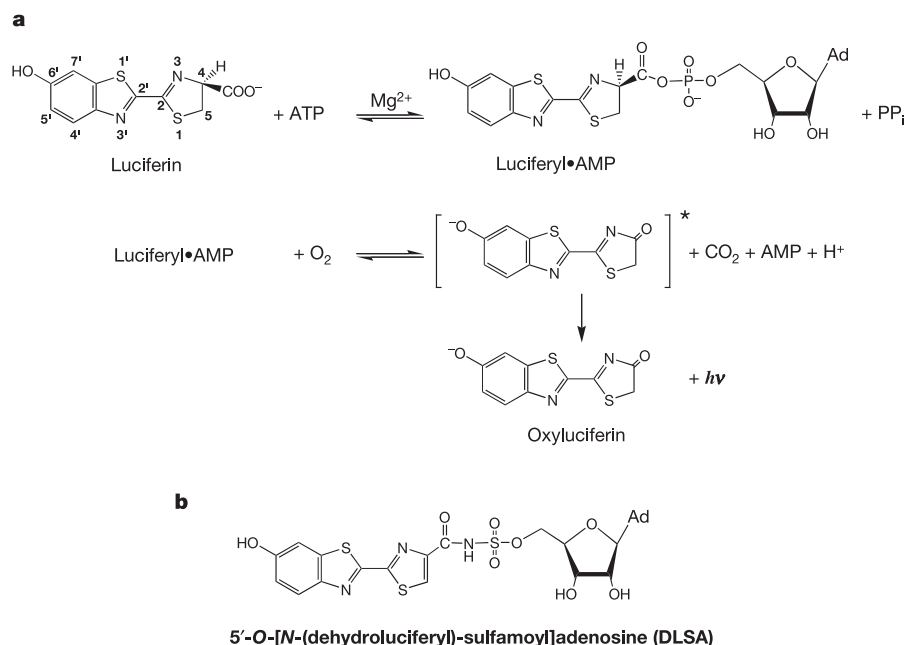


Figure 1 | The bioluminescence reaction catalysed by luciferase. a, A two-step reaction mechanism via the luciferyl-AMP intermediate. **b**, Structure of DLSA, a luciferyl-AMP intermediate analogue.

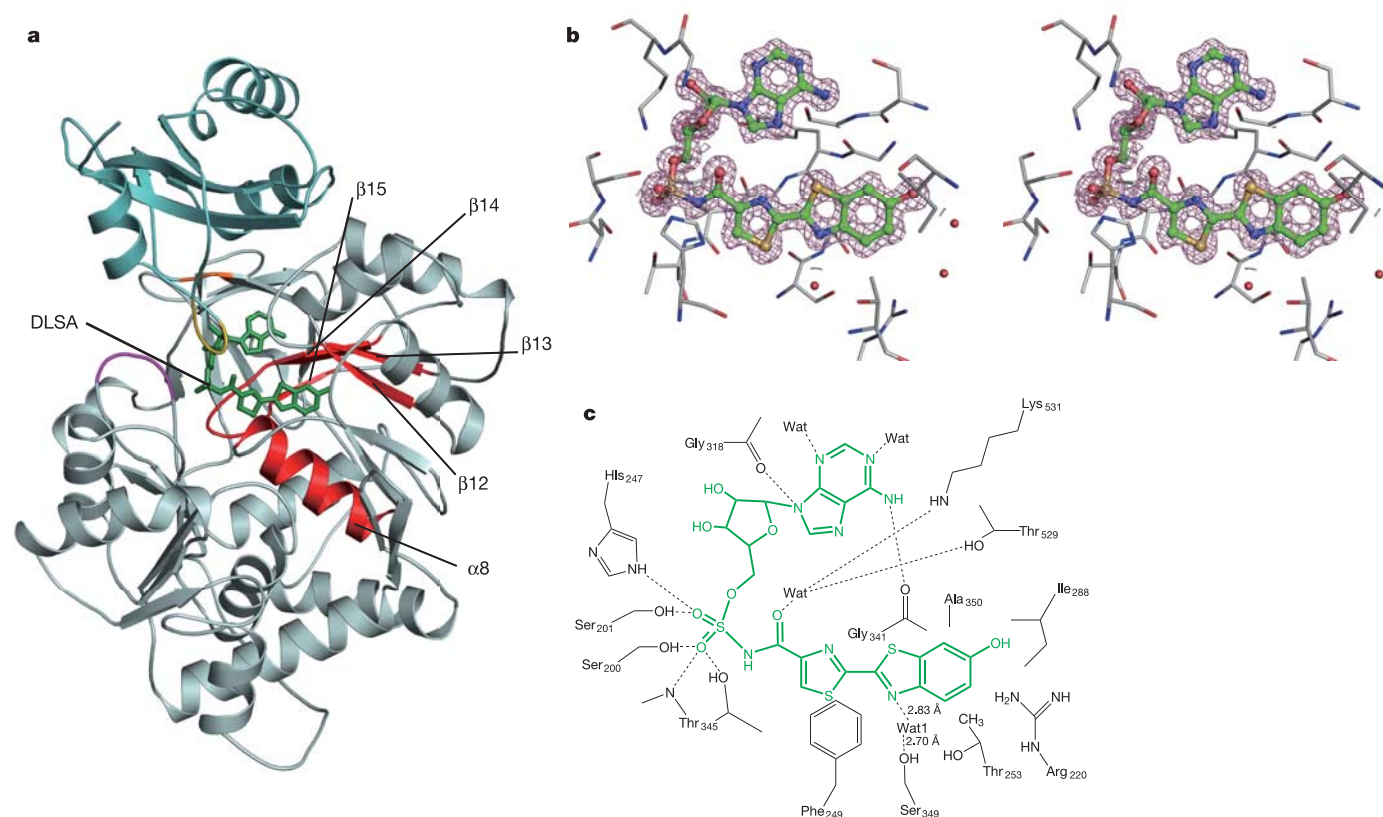


Figure 2 | Structure of Lcr luciferase-DLSA complex and DLSA binding site. a, Ribbon diagram of the LcrLuc(WT) in complex with DLSA (green). The N-terminal large and C-terminal small domains are drawn in grey and blue, respectively. The secondary structures for the luciferin-binding site are coloured in red. The P-loop¹⁷ (amino acids 201–205) is shown in magenta, the linker-loop (438–442) in orange and the active-site loop (526–530) in

yellow. These loops are not observed in the structure of *Photinus pyralis* luciferase. **b**, Stereo view of the structure around DLSA and $2F_o - F_c$ electron density map of DLSA contoured at 1.0σ . **c**, A schematic drawing of the DLSA binding site. DLSA is coloured in green. Polar interactions are shown as dashed lines.

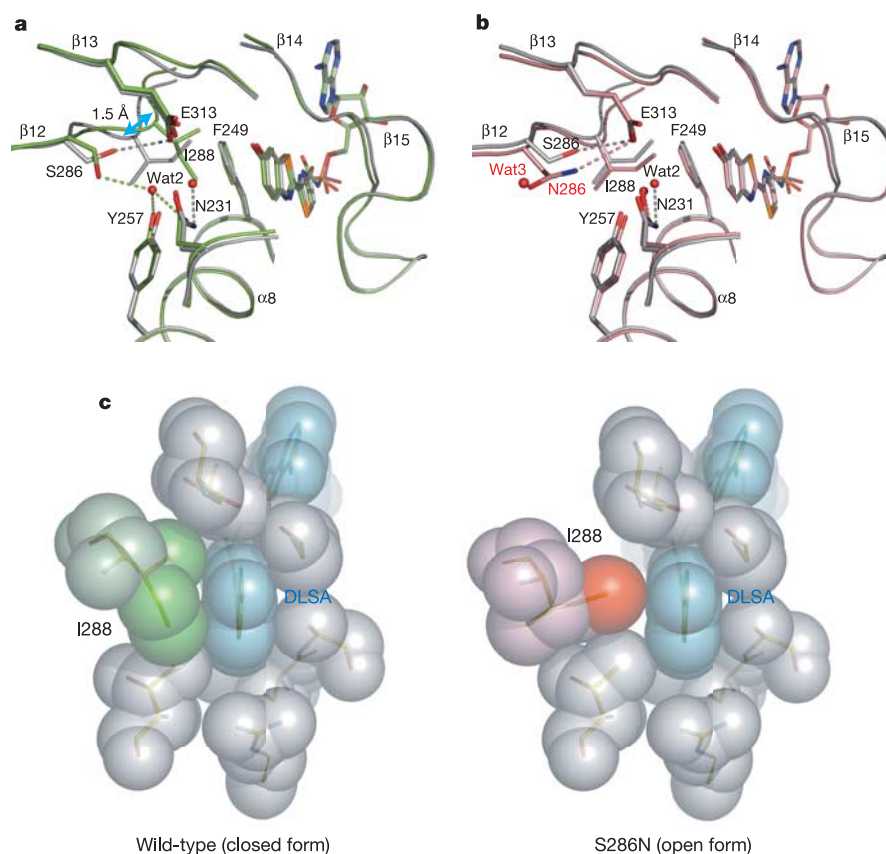


Figure 3 | Comparison of the luciferin binding site in the Lcr luciferase structures. **a**, Superposition of the LcrLuc(WT)·DLSA (green) and LcrLuc(WT)·AMP/oxyLuciferin (white) complexes. **b**, Superposition of the structures of LcrLuc(S286N)·DLSA (pink) and LcrLuc(WT)·AMP/oxyLuciferin (white) complexes. Polar interactions are shown as dashed lines. **c**, Comparison of van der Waals interactions in the structures of wild-type

(left) and S286N (right) luciferase complexed with DLSA. The van der Waals radii of DLSA (blue), Ile 288 of wild-type (green and light green) luciferase, and Ile 288 of S286N (red and pink) luciferase are drawn in colour. The atoms in van der Waals contact with DLSA are highlighted in green and red.

structure of the DLSA complex. Instead, Ser 286 in the DLSA complex forms new hydrogen bonds with Tyr 257 and Asn 231 via a water molecule (Wat 2) (Fig. 3a). This conformational change is also accompanied by rotation of the side chain of Phe 249 by 6° towards DLSA. These movements result in a 'closed form' of the active site, thereby creating a structure in which the benzothiazole ring of DLSA is tightly sandwiched in a hydrophobic pocket, with the side chains of Ile 288 and Phe 249 on one side and the side chain of Ala 350 and the main chain of Gly 341 on the other. Therefore, the wild-type enzyme changes its conformation during catalysis to make an extremely hydrophobic microenvironment for the excited state of oxyLuciferin by the transient movement of the side chains of Ile 288.

The single amino acid substitution Ser 286 to Asn changes the colour of the bioluminescence emitted by luciferase from yellow-green to red². We therefore determined the crystal structure of the red S286N mutant protein, LcrLuc(S286N), in complex with DLSA at 1.45 Å resolution, and compared the structure with the various wild-type enzyme structures. Notably, the conformational change observed in the wild-type enzyme was not seen in the S286N mutant. The structure of the S286N·DLSA complex does not adopt the closed form of the active site: the side chain of Ile 288 does not move closer to the benzothiazole ring of DLSA. This might be due to the Oδ1 and Nδ2 atoms of Asn 286 hydrogen-bonding to Glu 313 and Wat 3, respectively, which seems to prevent a structural change by Ile 288 (Fig. 3b). This is in sharp contrast to the structure of the wild-type enzyme in complex with DLSA (see Supplementary Fig. S1c). As a result, the structure of the S286N·DLSA complex near the Ile 288 residue is almost identical to that of the wild-type enzyme bound to AMP and oxyLuciferin (Fig. 3b). This structure, referred to as the

'open form' of the active site, is also observed with the wild-type enzyme in complex with Mg-ATP (Supplementary Fig. S1b).

As shown in Fig. 3c, the closed form of the active site in the LcrLuc(WT)·DLSA complex allows for close van der Waals contact between three atoms (Cγ1, Cγ2 and Cδ1) of the side chain of Ile 288 and the benzothiazole ring of DLSA, whereas the side chain of Ile 288 in the S286N mutant does not move towards DLSA, creating a less rigid and less hydrophobic microenvironment. The thiazole ring is almost coplanar to the benzothiazole ring for both the wild-type and the mutant enzyme, arguing against the view that spectral differences are caused by the rotational angle of the thiazoline/benzothiazole rings of oxyLuciferin¹¹. Therefore, the conformational change that controls the hydrophobic microenvironment for the excited state of oxyLuciferin is a key to understanding the bioluminescence colour. The open form of the active site represents the structure of luciferase in the ground state before and after the reaction, whereas the closed form corresponds to the active form of the enzyme during catalysis. The wild-type enzyme changes the conformation of its active site from the open to the closed form as the reaction proceeds to develop the luciferyl-AMP intermediate, and binds the excited state of oxyLuciferin tightly in a highly rigid and nonpolar microenvironment created by the hydrophobic side chain of Ile 288, thereby minimizing energy loss before emitting yellow-green light (see Supplementary Video S1).

In contrast, this conformational change is less favourable with the S286N mutant owing to its hydrogen-bonding network around Asn 286, and the conformation of the active site seems to remain in the open form during catalysis. The resulting, less rigid microenvironment allows some energy loss from the excited state, probably

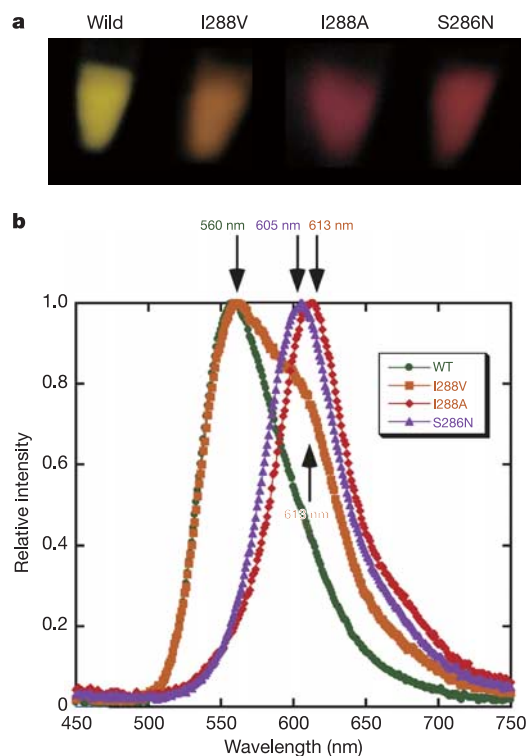


Figure 4 | Bioluminescence colour of wild-type and three mutant forms (I288V, I288A and S286N) of *Lcr* luciferases. **a**, Photographs of bioluminescence by a luciferase-catalysed reaction. The reaction condition was 0.1 mM luciferin, 2 mM ATP, 6 mM MgSO_4 , 25 mM HEPES (pH 7.8) and 0.1 mg ml^{-1} luciferase in 0.5 ml solution. **b**, Emission spectra of the four types of *Luciola cruciata* luciferases.

due to thermal relaxation or the reorganization of specific interactions, and leads to the emission of low-energy red light (see Supplementary Video S2).

To confirm this idea, we prepared the LcrLuc I288V and I288A mutants and examined the differences in their emission spectra. As expected, the I288V and I288A mutants emitted orange- and red-coloured light, respectively, and the extent of the spectral shift correlated with the size and hydrophobicity of the side chain (Fig. 4a). Therefore, the amino acid residue at position 288 directly influences the wavelength of the emitted light. Closer scrutiny of the emission spectra, however, revealed an interesting finding regarding the excited state. As shown in Fig. 4b, the wild-type enzyme, the I288A mutant and the S286N mutant gave single emission maxima at 560 nm (yellow-green), 613 nm (red) and 605 nm (red), respectively. On the other hand, two emission maxima (at 560 nm and 613 nm) were observed with the I288V mutant, suggesting that there are at least two different excited states that differ in energy level: one with a higher energy, emitting yellow-green light (560 nm), and the other with a lower energy level, emitting red light (613 nm). The excited states seem to shift from a higher to a lower energy level as the amino acid residue at position 288 changes from an Ile to Val to Ala (decreasing the size and the hydrophobicity of the side chain). The electronic structures of the excited states, however, have yet to be defined. Finally, the critical role of Ile 288 in the luminescence reaction has not been previously reported, but sequence alignment of a series of luciferases seems to highlight an integral function. Most firefly luciferases that emit yellow-green light have an Ile or a Leu residue at this amino acid position, whereas other luciferases that emit differently coloured light do not (see Supplementary Discussion).

METHODS

Crystallization. Details of mutational analysis, expression and purification are described in the Supplementary Methods. Luciferases (12–17 mg ml^{-1}) were

incubated at 20 °C for 16 h with ligands under three conditions: (1) 13.3 mM ATP and 26.9 mM MgCl_2 ; (2) 4.8 mM ATP, 1.6 mM D-luciferin and 16 mM MgCl_2 ; (3) 1.59 mM DLSA and 16 mM MgCl_2 . The incubated enzyme was filtered, and 3 μl of the filtered enzyme solution was mixed with an equal volume of reservoir solution (0.1 M Tris-chloride (pH 8.0–8.5) or 0.1 M HEPES-Na (pH 7.8), 20–23% PEG4000, 0.2 M lithium chloride, 10 mM 2-mercaptoethanol) to form a hanging drop. The drop was then equilibrated over 500 μl of reservoir solution at 20 °C. Clusters of crystals appeared after about one week. The spontaneous crystals were used for microseeding. The crystals obtained were flash-frozen in liquid nitrogen after soaking in a cryoprotectant solution consisting of the respective reservoir solution with a 2–3% higher PEG4000 concentration plus 20% (v/v) isopropanol, and used for X-ray diffraction experiments.

Data collection and structure determination. X-ray diffraction data of the LcrLuc(WT)-Mg-ATP complex crystal were collected on an R-Axis IIC with a Rigaku RU-300 rotating anode at room temperature (about 20 °C). The crystals, which belong to space group $P2_12_12_1$, with unit cell parameters $a = 57.8$ Å, $b = 182.0$ Å, $c = 53.6$ Å, diffracted to 2.3 Å resolution. X-ray diffraction data of LcrLuc(WT)-DLSA and LcrLuc(WT)-AMP/oxyluciferin crystals were collected at 90 K with wavelength of 1.02 Å on RIKEN Structural Biology Beamline I (BL45XU)²¹, and LcrLuc(S286N)-DLSA crystal was collected at 90 K with wavelength 1.0 Å on RIKEN Structural Biology Beamline II (BL44B2)²² at SPring-8. Data were processed with CrystalClear²³ (for the three types of LcrLuc(WT) crystals) or MOSFLM²⁴ and SCALA (for the LcrLuc(S286N)-DLSA crystal). Data collection statistics are shown in Supplementary Table S1.

Initial phases for the wild-type-Mg-ATP complex structures were obtained by molecular replacement using AMORE²⁵, with the N-terminal domain of luciferase from *Photinus pyralis* as a search model. Electron density maps of the N-terminal domains were obtained (the electron density maps of the C-terminal domains are not shown). After the N-terminal domain structure model was refined by Refmac²⁶, the interpretable sigmaA-weighted electron density map of the C-terminal domain was apparent and the C-terminal domain model was added. The overall structure model was refined by Refmac²⁶. The phases of the other complex structures were calculated using the LcrLuc(WT)-Mg-ATP complex structure and these crystallographic protein models were refined by Refmac²⁶ with ARP/wARP²⁷ automatic model building and water picking. Manual model building was performed using TURBO-FRODO²⁸. Structure determination and refinement statistics are shown in Supplementary Table S1. Molecular graphics were illustrated with PyMOL²⁹ (Figs 2a, b and 3).

Received 7 September; accepted 23 December 2005.

- Seliger, H. H. & McElroy, W. D. Spectral emission and quantum yield of firefly bioluminescence. *Arch. Biochem. Biophys.* **88**, 136–141 (1960).
- Kajiyama, N. & Nakano, E. Isolation and characterization of mutants of firefly luciferase which produce different colors of light. *Protein Eng.* **4**, 691–693 (1991).
- Mamaev, S. V., Laikhter, A. L., Arslan, T. & Hecht, S. M. Firefly luciferase: Alteration of the color of emitted light resulting from substitutions at position 286. *J. Am. Chem. Soc.* **118**, 7243–7244 (1996).
- Branchini, B. R., Southworth, T. L., Murtiashaw, M. H., Boije, H. & Fleet, S. E. A mutagenesis study of the putative luciferin binding site residues of firefly luciferase. *Biochemistry* **42**, 10429–10436 (2003).
- Branchini, B. R. *et al.* Site-directed mutagenesis of firefly luciferase active site amino acids: a proposed model for bioluminescence color. *Biochemistry* **38**, 13223–13230 (1999).
- Conti, E., Franks, N. P. & Brick, P. Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure* **4**, 287–298 (1996).
- Franks, N. P., Jenkins, A., Conti, E., Lieb, W. R. & Brick, P. Structural basis for the inhibition of firefly luciferase by a general anesthetic. *Biophys. J.* **75**, 2205–2211 (1998).
- DeLuca, M. Hydrophobic nature of the active site of firefly luciferase. *Biochemistry* **8**, 160–166 (1969).
- White, E. H., Rapaport, E., Hopkins, T. A. & Seliger, H. H. Chemi- and bioluminescence of firefly luciferin. *J. Am. Chem. Soc.* **91**, 2178–2180 (1969).
- Branchini, B. R. *et al.* An alternative mechanism of bioluminescence color determination in firefly luciferase. *Biochemistry* **43**, 7255–7262 (2004).
- McCapra, F., Gilfoyle, D. J., Young, D. W., Church, N. J. & Spencer, P. in *Bioluminescence and Chemiluminescence: Fundamental and Applied Aspects* (eds Campbell, A. K., Krick, L. J. & Stanley, P. E.) 387–391 (John Wiley and Sons, Chichester, 1994).
- DeLuca, M. Firefly luciferase. *Adv. Enzymol.* **44**, 37–68 (1976).
- Delarue, M. Aminoacyl-tRNA synthetases. *Curr. Opin. Struct. Biol.* **5**, 48–55 (1995).
- Chang, K. H., Xiang, H. & Dunaway-Mariano, D. Acyl-adenylate motif of the acyl-adenylate/thioester-forming enzyme superfamily: a site-directed mutagenesis study with the *Pseudomonas* sp. strain CB53 4-chlorobenzoate:coenzyme A ligase. *Biochemistry* **36**, 15650–15659 (1997).

15. Kleinkauf, H. & Von Dohren, H. A nonribosomal system of peptide biosynthesis. *Eur. J. Biochem.* **236**, 335–351 (1996).
16. Babbitt, P. C. *et al.* Ancestry of the 4-chlorobenzoate dehalogenase: analysis of amino acid sequence identities among families of acyl:adenyl ligases, enoyl-CoA hydratases/isomerases, and acyl-CoA thioesterases. *Biochemistry* **31**, 5594–5604 (1992).
17. May, J. J., Kessler, N., Marahiel, M. A. & Stubbs, M. T. Crystal structure of DhbE, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proc. Natl Acad. Sci. USA* **99**, 12120–12125 (2002).
18. Ueda, H. *et al.* X-ray crystallographic conformational study of 5'-O-[N-(L-alanyl)-sulfamoyl]adenosine, a substrate analogue for alanyl-tRNA synthetase. *Biochim. Biophys. Acta* **1080**, 126–134 (1991).
19. Branchini, B. R., Murtiashaw, M. H., Carmody, J. N., Mygatt, E. E. & Southworth, T. L. Synthesis of an N-acyl sulfamate analog of luciferyl-AMP: a stable and potent inhibitor of firefly luciferase. *Bioorg. Med. Chem. Lett.* **15**, 3860–3864 (2005).
20. Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.* **16**, 4174–4183 (1997).
21. Kumasaka, T., Yamamoto, M., Yamashita, E., Moriyama, H. & Ueki, T. Trichromatic concept optimizes MAD experiments in synchrotron X-ray crystallography. *Structure* **10**, 1205–1210 (2002).
22. Adachi, S. *et al.* The RIKEN structural biology beamline II (BL44B2) at the SPring-8. *Nucl. Instrum. Methods A* **467–468**, 711–714 (2001).
23. Pflugrath, J. W. The finer things in X-ray diffraction data collection. *Acta Crystallogr. D* **55**, 1718–1725 (1999).
24. Leslie, A. G. W. in *Joint CCP4+ESF-EAMCB Newsletter on Protein Crystallography* No. 26 (CCP4, Warrington, 1992).
25. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
26. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
27. Perrakis, A., Morris, R. & Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6**, 458–463 (1999).
28. Roussel, A. & Cambillau, C. *Turbo-Frodo Manual* (AFMB-CNRS, Marseille, 1996).
29. DeLano, W. L. The PyMOL Molecular Graphics System. (DeLano Scientific, San Carlos, California, 2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Kajiyama and K. Hirokawa (Kikkoman Corp.) for providing cDNA of luciferase and measurement of bioluminescence spectra. We also thank M. Yamamoto and S. Adachi (RIKEN Harima Institute) for X-ray diffraction data collection on the RIKEN beamlines BL45XU and BL44B2 at SPring-8. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan (T.N. and H.K.).

Author Information Atomic coordinates for luciferase structures have been deposited in the Protein Data Bank under accession codes 2D1Q (WT·Mg-ATP), 2D1S (WT·DLSA), 2D1R (WT·AMP/oxyluciferin) and 2D1T (S286N·DLSA). The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.K. (katohiro@pharm.kyoto-u.ac.jp).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Bench partners

Postdocs and grad students are often called the engine of the research enterprise. That may be true — up to a point. But it leaves out one group that is essential for powering the process: lab technicians.

Data on this group are hard to find: how many are there, what degrees do they hold, and what are their career trajectories? The traditional view is that technicians hold a bachelor's degree or equivalent, have a fixed skill set and often remain entrenched in the role.

But anecdotes that I have collected over the past few years run counter to that. Some students take technician jobs between undergraduate and graduate degrees. Technicians constantly learn new scientific skills and sometimes rise up the management ladder. One former technician I met at the University of Edinburgh, UK, took training courses and is now an administrator there. A columnist on the *Naturejobs* Graduate Journal took a few years off between degrees to work as an industrial technician; he has since gone on to complete a PhD and is now in the middle of a postdoc at the US National Institutes of Health. In one Scottish biotechnology company, a group of technicians completely

shifted the firm's scientific focus in six months. These examples belie the belief that technicians lack career mobility and seldom acquire fresh skills.

But there is more to be learned. What formal training programmes exist for career advancement? How many doctoral students who prefer research to grant writing opt to be technicians after their graduate work? Are technicians able to switch between sectors, and how many go on to excel in non-bench activity, such as management, business development, marketing and regulatory affairs? With the launch of the *Naturejobs* online technicians' portal we aim to explore those questions — because even if technicians are not the engine of research, they are certainly vital cogs.

NATUREJOBS TECHNICIAN FOCUS

► www.nature.com/naturejobs/focus/online-tech.html



Paul Smaglik, *Naturejobs* editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

European Head Office, London
 The Macmillan Building,
 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

European Sales Manager:
 Andy Douglas (4975)
Natureevents: Silke Opstrup (4994)

UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
**Germany/Austria/
 The Netherlands:**
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

**European Satellite Office
 Germany/Austria/
 The Netherlands:**
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 75 Varick Street,
 9th Floor, New York,

NY 10013-1917
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

CHEMISTRY'S EVOLUTION



N. GROSJEAN/CORBIS SYGMA

Zöe Schnepf may be green, but that's not a sign of inexperience. Named by the Royal Society of Chemistry as best chemistry student of 2005, Schnepf was at the forefront of changes in the discipline when she decided to focus her career efforts on green chemistry.

Green chemistry is a set of principles dedicated to creating more efficient industrial chemicals, drugs and products, driven by a mixture of political, economic and cultural factors. The economic drive is to reduce waste. The political drive comes from regulations, such as the US Pollution Protection Act, that are forcing companies to develop cleaner processes. And the cultural drive is provided by consumers and scientists who are becoming more aware of the need for cleaner processes.

Such factors helped to launch the field; recent recognitions may help legitimize it. The 2005 Nobel Prize in Chemistry was awarded to a group who developed a green organic-synthesis method used in the pharmaceutical and chemical industries. The award gives scientists such as Schnepf hope that a career can be made applying cleaner chemical approaches.

Another award, by the US Environmental Protection Agency (EPA), shows that green chemistry is more

Industry's need to reduce waste and deal with the environmental concerns of consumers is creating demand for cleaner catalysis, says Virginia Gewin.

than an industry buzzword. In the ten years the agency has presented the Green Chemistry Awards, the companies that won them have cut the amount of hazardous material or waste they produce by about 1.5 million tonnes. One recent winner, the drug company Merck, shows how greener can be cheaper. The EPA rewarded the company for changing its synthesis of Emend, an anti-nausea drug for chemotherapy patients, by cutting the number of steps to produce it from six to three and eliminating some hazardous compounds from the process. This reduced its cost and environmental impact.

Such approaches are attractive to all industries looking to cut costs, avoid lawsuits and maintain good public relations. So chemists interested in taking a green approach should find growing opportunities in chemical, drug and healthcare companies. But hiring managers are split over which is more desirable: specific green chemical training — a growing trend — or a more general background.

For hiring managers at the multinational General Electric (GE), green qualifications are less important than raw talent. "I don't go out and look for green chemists," says Greg Chambers, GE's global technology leader of polymer and chemical technology. His colleague Terry Leib, manager of GE's material-analysis and chemical-sciences division, says green chemists have some advantages, including greater awareness of environmental issues.

The biggest potential growth area in green chemistry is in finding cleaner, cheaper catalysts — because the chemicals used to accelerate reactions for everything from gas to detergents can also cause waste and pollution. Developing catalysts that are in a different phase from the reaction species — a solid among liquid reactants, for example — simplifies the purification of the end product, reducing the amount of energy used and waste produced. And chemists are working on such heterogeneous catalysts to ensure they can be easily recovered and reused. Just last year, researchers at Iowa State University developed a low-cost solid catalyst that can be recycled more than 20 times without any drop in performance.

Basic and applied: companies such as DuPont (above) are adopting green chemistry. And Richard Schrock (left), Yves Chauvin (centre) and Robert Grubbs shared a Nobel last year for their work on environmentally friendly catalysis.



AP PHOTO/J. EKSTROMER

Mark Harmer, senior scientist at DuPont's central research and development department, emphasizes that green chemistry has industrial economics on its side. Developing heterogeneous catalysts can reduce the amount of waste generated during product development by 95–99%. "It takes time and effort but companies have to commit to looking over the longer term," he says.

As part of a \$1.5-billion eco-imagination programme, GE plans to double its spending on developing cleaner technologies and reducing greenhouse-gas emissions by 2010, through projects such as hybrid locomotives. The quest to reduce waste and comply with regulations is driving other companies to recruit green chemists. Colja Laane, corporate science manager at chemical manufacturer DSM in Heerlen, the Netherlands, expects jobs to increase as chemical industries invest in innovation. DSM, for example, will hire an extra 250 people in areas from research to engineering, processing and business development, to support its €75-million (US\$89-million) investment in innovation, which focuses heavily on using living cells to produce enzymes for use as catalysts.

As well as chemical manufacturers, makers of consumer goods are increasingly focused on their environmental impact. S. C. Johnson, which makes cleaning products, developed Greenlist, an in-house three-point system to classify raw materials used in manufacturing based on their impact on the environment and human health. "We use environmental toxicologists to look at profiles of a product from environmental and human standpoints and choose those chemicals that have least risk," says Scott Johnson, the company's vice-president of global environmental and safety actions. Over the past five years, he adds, they have made steady progress in reducing their environmental impact.

Two of the largest drug companies, Pfizer and GlaxoSmithKline, have used green chemistry to reduce waste. Pfizer has used green techniques to decrease the amount of waste generated for each kilogram of drug produced from from 25–100 kg to 10–25 kg. "We see green chemistry not as a revolution, but an evolution," says Peter Dunn, Pfizer's head of green chemistry. The company's focus on sustainability has placed an emphasis on quickly identifying cheap green solutions. In addition to chemical technologists trained in high-throughput screening that will find greener reagents or solvents, those trained in biotransformations, or shuffling genes into a water-based enzyme that generates less waste, are in demand. The quest for



From green to gold? Steve Howdle (top, right) and Robin Rogers (above, left) have each launched green chemistry start-ups.

WEB LINKS

Green Chemistry Institute
 ◆ tinyurl.com/acqbt
 Green Chemistry Network
 ◆ www.chemsoc.org/networks/gcn
 SusChem
 ◆ www.suschem.org
 Center for Environmentally Beneficial Catalysis
 ◆ www.ku.edu/~cebc
 Carnegie Mellon Institute for Green Oxidation Chemistry
 ◆ www.chem.cmu.edu/groups/collins
 University of Massachusetts at Lowell programme
 ◆ www.greenchemistry.uml.edu
 University of York programme
 ◆ www.york.ac.uk/res/gcg

innovation is even fuelling an entrepreneurial spirit. Robin Rogers, a chemist at the University of Alabama in Tuscaloosa, points out that green chemistry, like biotechnology before it, will be most innovative in the arena of start-ups. "The best, most profitable technologies will be sold to or adopted by major industries," he says. To that end, he has created a start-up called 525Solutions to develop alternatives to traditional solvents. Steve Howdle, a chemist at the University of Nottingham, UK, researches polymers that will generate less pollution during the manufacturing process. One of his ideas, a biodegradable polymer that releases drugs within the body, is the focus of a company he is starting.

A handful of green programmes are training chemists to think specifically about cleaner synthesis. John Warner, director of the green-chemistry centre at the University of Massachusetts at Lowell, puts mastery of traditional chemistry skills above the green label, but he says that green-chemistry students are uniquely positioned to address industry concerns because of their specific training in both industry regulations and manufacturing constraints. His colleague, Terry Collins, director of Carnegie Mellon's Institute for Green Oxidation Chemistry, agrees with him. Collins invented tetraamido macrocyclic ligand (TAML) iron (III) activators — synthetic catalysts that significantly increase the oxidizing ability of hydrogen peroxide. Compared with previous techniques, much smaller quantities of TAML activators are needed to completely degrade harmful chlorinated organics, such as dioxins. "It's a dream solution to a 50-year old problem," Collins says, adding that the compounds can also clean up problematic organophosphates.

Growing industry interest in greener catalysis methods helped to form the Center for Environmentally Beneficial Catalysis (CEBC), a National Science Foundation Engineering Research Center at the University of Kansas in Lawrence. Not surprisingly, Bala Subramaniam, the CEBC's director, says the centre's 60 students are highly sought after by industry members, including DuPont, ExxonMobil and the agricultural processor Archer Daniels Midland of Decatur, Illinois.

The first generation of green chemists are busily training the next generation, says Collins, and students are getting snapped up. "Of the 100 students to pass through my research lab, none has ever gone longer than three weeks looking for a job," says Warner. One graduate, Amy Cannon — now a consultant for the materials company Rohm and Haas — says that although her title of 'green chemist' raises initial scepticism from those in industry, those same industry leaders inevitably seek her advice.

"The biggest challenge is that there are more jobs out there for green chemists than there are trained chemists for the jobs," says Paul Anastas, director of the Green Chemistry Institute in Washington DC and coiner of the term. He is borne out by a 2005 US National Research Council report that identifies a need to increase training in green chemistry — to spur the move away from fossil fuels over coming decades. The report recommends incorporating green-chemistry principles into education materials for students as well as business leaders.

Virginia Gewin is a freelance science writer based in Portland, Oregon

EUROPE'S CHEMICAL WASTAGE

Worldwide, chemistry has not been a popular career choice in recent years. Indeed, a steady decline of chemistry students in Europe has prompted concern. With 1.7 million people employed in the chemical industry in the 25 countries of the European Union, the industry is fighting to remain a competitive employer.

One programme, the European Technology Platform for Sustainable Chemistry, or SusChem, is promoting a research agenda emphasizing cleaner industrial technologies. "If we don't do something, there will be radical decline in industry employment," says Russell Mills of Dow Chemical, who sits on SusChem's board. Expertise in biocatalysis, process design and nanotechnology are some of the future skills that will be critical. To increase work in these areas, SusChem hopes to boost the European Union's funding of training and research in chemistry by 75%, to €5.5 billion (US\$6.6 billion) by 2013.

V.G.

MOVERS

Andrew Chien, director of research and vice-president of corporate technology group Intel, based in Hillsboro, Oregon



2003-05: Founding director, Center for Networked Systems, University of California, San Diego

1998-2005 Professor of computer science, University of California, San Diego

1999-2003 Co-founder, chief technology officer and chairman of the board, Entropia, San Diego, California

As an undergraduate at the Massachusetts Institute of Technology (MIT), Andrew Chien thought he would become a doctor. When a glass-blowing seminar he was taking for fun was cancelled, he followed a friend to check out the "crazy engineering teacher" — Gerald Sussman, a computer programmer legendary for his enthusiasm. Chien was thenceforth hooked on computers and engineering.

Corporate giants and entrepreneurs were always Chien's heroes. But near the end of his doctorate at MIT, his adviser, Bill Dally, though a successful entrepreneur himself, convinced him to spend time in academia first. Chien took a junior faculty position at the University of Illinois at Urbana-Champaign, where his programming and computer-architecture work set him up for early tenure. If that wasn't enough to cause a stir on campus, his request to do his sabbatical in an industrial research lab did. He convinced his dean to let him spend a year at the Hewlett-Packard lab in Palo Alto, California, where he got his first taste of the corporate research environment.

At the time, California was a hot spot for computing and engineering. When the University of California, San Diego (UCSD) offered him an endowed chair — at the age of 34 — he accepted it to take advantage of the area's wealth of industrial collaboration and start-up opportunities in computing, networking and bioinformatics.

He was finally able to indulge his inner chief executive at the helm of an information technology start-up called Entropia in San Diego. It fell victim to the bursting technology bubble in 2000 but he has since straddled the line between academia and industry, being involved with a number of leading companies.

"Academia's appeal is the tremendous freedom to follow research interests, and industry's advantage is being much closer to the innovations that transform society," he says.

Three years ago, he created a multipartner academic-industry research centre at UCSD called the Center for Networked Systems. Setting mutually beneficial technical agendas while juggling intellectual-property issues and long-term financial commitments prepared him for his new position as director of Intel Research. His greatest challenge yet, he says, will be to drive the company's research agenda. He hopes to get more people hooked.

"I really encourage scientists to spend time in corporate America," he says. "There's much to be learned that can make you a stronger scientist overall."

Virginia Gewin

RECRUITERS & ACADEMIA

What's holding you back?

Through hundreds of consultations with young researchers, I have heard many say that they want to "get away from the bench" while still wanting to "use their science". Identifying how they can do this is challenging and requires recognition of both the practical and psychological forces at play.

Practical career advice is widely available. But the psychological dimensions, equally significant, are less commonly addressed and understood. Those willing to modify their attitudes to themselves and their profession can develop highly satisfying careers. Here are three common psychological pitfalls that impede young scientists.

Blood from a stone. It can be tempting to believe that one's specific area of scientific knowledge will be the source of the best job opportunities. But the greatest value from a PhD or post-doctoral training experience will often be the knowledge gained of research methods, the acquisition of analytical skills, the ability to communicate with and teach a variety of audiences, and to interpret science at the highest level — all can be valuable outside the lab. A belief that doing something other than research is tantamount to leaving science can hold people back from using their knowledge and skills in other meaningful and rewarding ways. Scientists who leverage their skills in

other fields tend to see themselves as adding to, not replacing, their identity.

The 'shining star' syndrome. Before arriving at graduate school, research trainees have typically stood out academically. Now, surrounded by other stand-outs, they may not feel so gifted. To strive for excellence is important, and I'm not advocating a celebration of mediocrity, but there is tremendous value in performing at a level sufficient to do what is needed. Humility is about as rare as that monumental scientific breakthrough, but is no less important.

Can't please everyone. Some young researchers see only a few career options as acceptable, based on parental, mentor and societal influences. Disappointing a parent or mentor can be traumatic, but can also allow one to discover self-responsibility and freedom of identity, a critical task. There comes a time when it's crucial to take steps towards defining, on one's own terms, career objectives and a means for reaching them.

Those who understand these barriers can work to overcome them and take the first step towards defining the careers and lives they truly want.

Michael Alvarez is founding director of the Stanford University School of Medicine Career Center.

♦ <http://med.stanford.edu/careercenter>

GRADUATE JOURNAL

Master of multitasking

It isn't until the end of graduate school that you begin to wonder whether you've acquired any transferable skills. Despite having been at school since the age of five, it suddenly occurs to you that you're not generally useful to the world.

I know what transferable skills I don't have. I once sat next to a hypochondriac on a plane who berated me for four hours for being a biologist but not being able to cure his adult-onset diabetes. Thing is, unless you are a yeast cell I can't really help you with your health problems. I can measure sporulation efficiency pretty well, but most humans don't actually turn into spores, it turns out.

Then I started watching the Olympics. The commentators kept telling me how much pressure these athletes are under. That's when it hit me. This month I taught a class, wrote an abstract for a thesis committee meeting, finished key experiments for two projects, caught up with the literature, met my adviser a few times, and in my spare time even made it to the supermarket once or twice. This would have seemed like a bit much to handle in my first year of graduate school. Now, it's par for the course. I'm not sure it's enough to get hired, but where were those commentators when I trudged to my lab during a blizzard because an experiment couldn't wait? Good at multitasking under pressure — that's one transferable skill. Now if only I could land a triple axle.

Milan de Vries is a molecular-biology graduate student at the Massachusetts Institute of Technology in Cambridge, Massachusetts.

Product development

Switch off, tune out, and clean up.

Nancy Kress

"That's the stupidest idea I've ever seen," the vice-president for marketing said, but so softly that only his neighbour heard. "What is the Old Man thinking?"

"Dunno," VP Sales whispered back. "But he loves it. Look at him."

The chief executive of Veritronics Telecommunications smiled at the head of the vast teak table in the vast corporate boardroom. Beside him fidgeted the head of R&D, a small pinch-faced man with the glowing maniacal eyes of a feverish gerbil. On R&D's palm rested a black plastic cube, featureless except for a simple red toggle switch.

"This model affects an area of radius 20 feet," R&D said fervently, "but we plan to create a whole range of models to cover homes and facilities of different sizes, maybe even different shapes. It'll make Veritronics a fortune!"

"Demonstrate it for them, Lucius," the CEO said.

R&D toggled the switch. Instantly the wall panels, which had been displaying a fractal composition by revered holo-artist Cameron Mbutu, went dark. Every handheld in the room stopped functioning, marooning VP Accounting, who'd been surreptitiously playing *Alien Attack*, on level 184. All personal receivers ceased operation, cutting off VP Sales from the field reports reciting softly in her ear; VP Admin from the London Philharmonic's performance of Haydn's Symphony No. 104; and Veritronics' General Counsel from the weather report for Cancun, where he was going on holiday. All cell phones stopped vibrating in all pockets. In the middle of the immense table, the electronic news screen blanked, along with the second-by-second stock reports from six cities and the lunch menu. Only the lights and heat stayed on.

VP Marketing and VP Sales glanced at each other. VP Sales was braver.

"Sir...why do you think anyone would want to jam their own home telecommunications? And — with all due respect, sir — why would we want them to?"

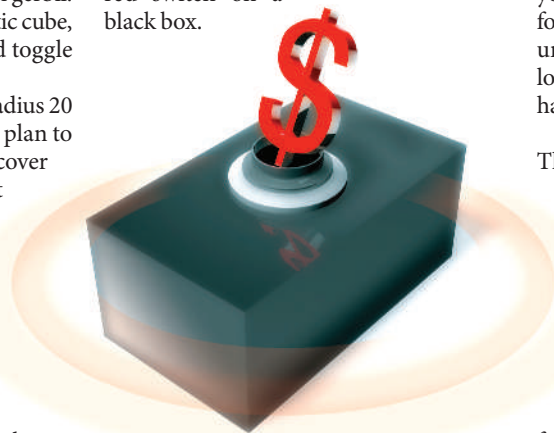
The CEO smiled. "Lucius, show them the tape."

R&D did something under the table. A single wall panel glowed. "This is a composite of the data from 146 beta-test trials. It has no margin of error."

A middle-class living room. "Jimmy!" a harried woman screeched. "Did you do your homework? Jimmy! Alia, I told you to watch the baby! Paul, I need help here!"

Jimmy, oblivious, beat time on the sofa arm to his personal receiver. Alia hunched over a video game, her fingers flying. Paul spoke rapidly into his handheld. The baby tumbled down a short flight of steps, its wailing nearly lost in the TV blare.

The woman snatched up the screaming baby, glared at her family, and reached into her pocket. CLOSE-UP as she toggled a red switch on a black box.



"Hey!" Jimmy cried. Alia continued to work her dead controls. Paul jumped up wildly. "Mia, what happened?"

"I don't know," Mia said innocently.

"Must be a power outage."

"The lights are still on!"

"Well...I don't know. I'll get cards and make popcorn, okay?" She smiled quietly.

VP Sales whispered: "Told you so. The Old Man's lost it."

VP Marketing didn't reply. Sweat banded his forehead and he clenched his hands tightly below the table.

The screen flashed TWO HOURS LATER. The family sat slumped over a card game. Paul snapped: "Mia, I told you and told you to not trump my ace."

"Well, I'm sorry! If you didn't always ... Jimmy, stop kicking me!"

"I can't help it," Jimmy whined.

"You know he's ADD," Paul said in disgust. "Why are you always riding him?"

"If you'd ever discipline him to..."

"I'm bored," Alia said. "And I'm s'posed to call Tara! It's important!"

"Tara can wait," Paul said.

"You never let me do anything! I wish I had different parents!"

Mia looked stricken. Paul said heatedly,

"You liked us well enough when we all went on that VR safari last month!"

"Yeah...but that was fun. Jimmy, stop kicking me!"

Jimmy threw his cards at Alia. Paul glared at Mia. "Great idea this was, genius!"

"Don't start with me, Paul. I'm as smart as you are even if not everybody can go to Harvard."

Alia burst into tears, waking the baby. "Don't start all that fighting again! Why don't you two just get divorced! I hate you!" She stomped from the room. Paul followed, slamming the door. Jimmy slunk under the table, kicking its legs. Mia looked around helplessly, then slipped her hand into her pocket.

"That's enough, Lucius," the CEO said. The wall went dark. "This family lasted two hours and three minutes before all their sublimated resentments and group tensions broke out. That's 16 minutes longer than average."

"But..." VP Sales began, then stopped. Her heart beat too hard and her left temple twitched.

"Within the next five hours," the CEO continued, "this family purchased four new electronic products, two of them from Veritronics."

VP Marketing dug the nails of one hand into the flesh of the other. Sweat slimed his forehead: "No...no margin of error, sir?"

"None. Every single test case exhibited the same behaviour."

"The same," croaked R&D. His foot beat a ragged staccato on the floor.

"Four new sales in five hours," General Counsel repeated. His face had paled to the unhealthy colour of sourdough.

The CEO toggled the red switch. Instantly the wall panels shone with gorgeous art. The table screens resumed their ceaseless supply of data. Haydn, *Alien Attack* and field reports all played. Cell phones vibrated. Handhelds glowed. There were 23 lunch options.

VP Marketing felt his breathing steady, his heart slow, his sweat evaporate. "I propose accelerated development and launch of the Veritronics Home Electronic Jamming Saviour, sir. Put it on immediate fast track!"

The vote was unanimous. ■

Nancy Kress's most recent novel is *Nothing Human*, in which the human race comes to an end — but not before engineering our genetic descendants. Her fiction has won three Nebula Awards and a Hugo Award.